MODELING REPEATED MULTIVARIATE DATA TO ESTIMATE INDIVIDUALS' TRAJECTORIES, AND RISKS OF MAJOR CLINICAL EVENTS WITH APPLICATION TO SCLERODERMA

by

Ji Soo Kim

A dissertation submitted to The Johns Hopkins University

in conformity with the requirements for the degree of

Doctor of Philosophy

Baltimore, Maryland

October, 2020

Thesis Committee

Primary Readers

Scott L. Zeger (Primary Advisor)
Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Ami A. Shah
Associate Professor
Division of Rheumatology, Department of Medicine
Johns Hopkins University School of Medicine

Abhirup Datta
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Michelle C. Carlson
Professor
Department of Mental Health
Johns Hopkins Bloomberg School of Public Health

Alternate Readers

Joanne Katz
Professor
Department of International Health
Johns Hopkins Bloomberg School of Public Health

Ni Zhao
Assistant Professor
Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

Abstract

A patient's health is reflected in multiple longitudinal biomarkers in many chronic diseases. To treat a patient, physicians need to integrate information across multiple parameters and organ systems, factoring in the patient's prior trajectory and baseline risk factors to estimate his/her current and future health state depending on treatments. Aggregating this complex, longitudinal data for clinical use requires a major time investment on the part of the treating physician when time with patients is short. It is also challenging to clearly explain this information to patients during a routine clinical visit to facilitate shared decision making.

In this thesis, we introduce a statistical framework and a set of data science tools we developed to facilitate and improve clinical care in information rich settings. Motivated by a case study of scleroderma, a complicated and heterogeneous disease that affects multiple organ systems, we build methods to help understand each individual's disease progression throughout time in multiple dimensions.

Modeling trajectories in multiple dimensions involves a choice of whether to jointly model all markers in a single model or to model each marker separately. We investigate the advantages and disadvantages of jointly modeling

the outcome variables instead of using separated models when both address the main scientific or clinical question. We present general formulae for the relative efficiency of the two model estimators and examine in detail the implications of these formulae on the scleroderma clinical data.

Then, we extend the framework developed to estimate individual trajectory to predicting patients' risk of having clinically defined critical events in the near future. We introduce a cross-validated sequential prediction (CVSP) algorithm that quantifies patients' risk of multiple important clinical events by predicting their future trajectories given the prior trajectory in multidimension and baseline risk factors.

Finally, we construct a web-based application that shows a patient's longitudinal data in multiple organ systems visualized against those of other similar patients in selected clinical subgroups. To improve patient care in clinical settings, we introduce our approach of implementing and testing the utility of the interactive interface to communicate visualizations of a patient's and reference population's longitudinal data.

Acknowledgments

Every moment in my PhD studies, I was thankful for being in an environment where I could learn something from everyone I encountered. I had the greatest privilege of having a brilliant and thoughtful advisor, Dr. Scott Zeger. Scott guided me to become a better statistician, researcher, and communicator, and more than anything I am grateful he made the learning process enjoyable. I was fortunate to have Dr. Ami Shah as my mentor, who introduced me to the fascinating world of Rheumatology. Throughout our collaboration, I was continuously amazed by her bright mind, leadership, and kindness. I would like to thank other collaborators at the Johns Hopkins Scleroderma Center, especially Dr. Laura Hummers.

I would like to thank Drs. Abhirup Datta and Michelle Carlson for serving on my thesis committee, and Drs. Joanne Katz and Ni Zhao for being my alternative examiners. I am thankful for all the wonderful faculty members in the Biostatistics department, all of whom have taught me so much. Special thanks to Drs. Ciprian Crainiceanu and John Muschelli for their support and guidance.

I am grateful to each and every member of the Zeger Lab for their support, feedback, and inspirational ideas. I would like to express my gratitude

especially to Drs. Yizhen Xu and Zhenke Wu for sharing their knowledge and insight. I appreciation also goes to thank Emily Scott and Dr. Shannon Wongvibulsin and for their help and friendship over the years.

I would also like to acknowledge fellow students I met at the School of Public Health who were amazing friends and colleagues. I would like to express my sincere gratitude especially to Dasom Jang and Lauren Lan; your friendship was one of the greatest joys of my PhD years. I am also deeply thankful to have been on this doctoral journey with Sophie Berube, Sara Wang, and Dr. Bonnie Smith.

My heartfelt appreciation goes to my parents, Drs. Hagki Kim and Soonae Kim, for their unconditional love, encouragement, and the opportunities they have given me. My thanks also go to my two brothers, Jiwoong Kim and Jongchan Kim, for their love and friendship.

I am grateful to have my dog Hodoo for the joy and comfort he has brought into my PhD life. Lastly, I would like to express my gratitude to my boyfriend Sean Yoon. Thank you for all your love and support and always cheering me on. I couldn't have done it without you.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

We live in an era marked by a plethora of data. Modern technologies in biology, information, and communication combined with recent scientific discoveries has generated vast amounts of novel data that relates to quantifying an individual's health including, for example, step counts, vital signs, and DNA sequences. In clinic, it is typical that the disease of interest is monitored using many biomarkers. Longitudinal clinical measures and occurrences of clinical events throughout a patient's course of disease represent opportunities to practice evidence-based clinical care but also pose significant challenges. Physicians need to integrate information across multiple parameters and organ systems, factoring in a patient's prior trajectory and baseline risk factors to estimate his/her current and future health state depending on treatments. Aggregating this complex, longitudinal data for clinical use requires a major time investment on the part of the treating physician when time with patients is short. It is also challenging to clearly explain this information to patients during a routine clinical visit to facilitate shared decision making.

This thesis introduces a statistical framework and a set of data science

tools we developed to facilitate and improve clinical care in information rich settings. Motivated by a case study of scleroderma, a complicated and heterogeneous disease that affects multiple organ systems, we build methods to help understand each individual's disease progression throughout time in multiple dimensions. The Johns Hopkins Scleroderma Center dataset has characteristics that are common in many problems where the individual unit's trajectory is the focus. We observe a mix of continuous measures and discrete events at irregularly observed times. Of the five continuous outcomes of interest, some clearly follow non-Gaussian distributions with means and variances depending upon predictor variables. We also have multiple predictors clinically known to influence the outcomes that need to be considered. In this thesis, we focus on building models that flexibly describe the population and individual's trajectories that can accommodate the heterogeneity of measures, observations and individuals.

Modeling trajectories in multiple dimensions involves a choice of whether to jointly model all markers in a single model or to model each marker separately. Either approach can be used to answer questions about within-measure contrasts, but the marker-specific models are not sufficient to answer contrasts across multiple measures. Because the joint model quantifies the covariation across measures, it can produce more valid estimates and inferences about cross-measure contrasts.

In problems of the first kind, when a "separated" set of models, fit one at a time, addresses the same question as a joint or "combined" model, the question remains: what are the relative merits of each approach? Separated

models are simpler to specify and fit because each measure can be handled in a univariate regression using standard software for model fitting, checking, and inferences. But using separated models implicitly assumes there is no correlation among the multiple measures at the same or different times. As discussed in the thesis, this assumption, when wrong, will result in a loss of efficiency.

In Chapter 2 of this thesis, we investigate the advantages and disadvantages of jointly modeling covariates instead of using separated models when both address the main scientific or clinical question. We present general formulae for the relative efficiency of the two model estimators and examine in detail the implications of these formulae on the motivating scleroderma clinical data. Our approach is somewhat unique in that we focus on relative performance of the two models on individual patient trajectories rather than the average trajectories for subpopulations.

The framework developed to estimate individual trajectory by jointly modeling multiple markers can be extended to predicting patients' risk of having critical events in the near future. For clinical care of many chronic diseases, it is often of importance to generate real-time, actionable predictions for sentinel events. For scleroderma patients, cardiomyopathy, pulmonary arterial hypertension (PAH), and interstitial lung disease (ILD) can result in significant morbidity and mortality, and early detection and therapy can improve clinical outcomes. Timely risk predictions are essential because they: (1) warn clinicians of higher risk in need of increased monitoring and interventions; (2) reduce concerns in patients at lower risk. Increased concern

for these events are defined by clinicians to be the crossing of biomarker thresholds. In chapter 3, we introduce a prediction algorithm that quantifies patients' risk of multiple important clinical events by predicting their future trajectories given the prior trajectory and baseline risk factors. The models used in prediction jointly fit multiple biomarkers as outcomes. This implies that, even for patients who lack data in one of the measures due to short follow-up time, we can expect better prediction compared to using a marker-specific model as we are borrowing strength from the information seen in other measures. We investigate the effect of jointly fitting multiple longitudinal outcomes and further evaluate our model by comparing to empirical models that directly models the critical events as the outcome. Another important feature of the prediction model is that it generates real-time prediction of risk as new data are observed without re-fitting the model.

The aforementioned methods are developed considering their use in the clinic to communicate a person's likely disease status, past trajectory, and predictions of what is expected in the coming period for scleroderma patients. In the current practice of medicine, a clinician has access to historical and current data only about the patient at hand. That information is not typically organized or presented in a fashion for the clinician to readily appreciate the current status relative to its past. In addition, the patient's data are not placed within the context of other similar patients. For example, outcomes for prior similar patients are not typically available. Clinicians therefore are forced to make qualitative judgements about the patient's status, trajectory, and likely benefits of different treatments, not fully informed by either the patient's own

data or the experiences for other similar patients.

In order to improve patient care, we construct a web-based application that shows a patient's longitudinal data in multiple organ systems visualized against those of other similar patients in selected clinical subgroups. Chapter 4 presents the design of the visualization tool, which provides the aggregate clinical phenotype and longitudinal data in a snapshot view. The tool includes cumulative disease manifestations, autoantibody status, and medication history among many other clinically relevant parameters. Our ultimate goal is to embed the tool within the clinical workflow used by physicians to guide their interactions with patients, thereby improving shared medical decision making. To do so, we must implement and test the utility of the interactive interface to communicate visualizations of a patient's and reference population's longitudinal data.

We further propose to study the value of this tool in the clinic by conducting a randomized clinical trial. The trial is designed to assess the usability and shared decision making of the tool from the designer, provider and patient perspective. Details of the evaluation of the tool are presented in Chapter 4. Once our visualization tools are implemented at Johns Hopkins and approved for wider use based upon the clinical trial designed here, we will enhance the tool by including predictions of future trajectories and the risks of organ-specific complications as described in the previous chapter. We will use a similar protocol to test the utility of the enhancements. We will then seek to disseminate this tool beyond Johns Hopkins for broader use in the rheumatology community.

Chapter 2

Modeling repeated multivariate data to estimate individuals' trajectories with application to scleroderma

2.1 Introduction

Scleroderma or systemic sclerosis is an autoimmune disease characterized by dysregulation of the immune system, vasculopathy and fibrosis of multiple organ systems, including the skin, heart, lungs, kidneys, gastrointestinal tract, and blood vessels (Pattanaik, Brown, and Postlethwaite, 2011). In the United States, scleroderma has annual incidence of 19.3 new cases per million adults and prevalence of 276 cases per million adults (Mayes et al., 2003). Although the disease is uncommon, scleroderma is a one of 80 related autoimmune diseases that, in aggregate, comprise the 3rd most prevalent set of chronic diseases after cancer and heart disease (Fairweather, Frisancho-Kiss, and Rose, 2008). Scleroderma patients suffer from high morbidity and mortality

with a median survival of 11 years (Mayes et al., 2003; Denton and Khanna, 2017). Mechanisms that cause the pathology still remain unknown but are thought to be common to many other autoimmune diseases with higher aggregate prevalence, and treatment is customized to the individual organ systems involved in a given patient (Steen, 2008; Shah and Wigley, 2013). Moreover, there is a wide heterogeneity among patient populations in terms of clinical manifestations, response to treatment, rate of disease progression, and survival, which adds another layer of complexity in understanding and treating scleroderma (Allanore et al., 2015).

Given the substantial heterogeneity within the disease, clinicians and scientists have been working to identify more homogenous patient subgroups that are likely to share underlying disease mechanisms and treatment strategies. Since the first description of scleroderma in 1842, understanding of the disease has advanced through finding and analyzing subgroups defined by the extent of skin involvement, lung involvement, or autoantibody profiles (Steen, 2008). Identifying such subgroups is essential to understanding the etiology of the disease and consequently to clinical treatments since it is known that patients of the same clinical or demographic subtype tend to share similar prognosis (Shah and Wigley, 2013; Denton and Khanna, 2017). Hence, it is crucial to identify patients with rapid progression or at high risk of organ failure and the associated biomarkers such as patterns in autoantibody profile.

To accurately measure patients' health state and rate of progression at a given moment, we present a method of optimally defining population and individuals' health trajectories. For many chronic diseases, patients' health

state is reflected in longitudinal clinical measures and/or occurrences of clinical events over their course of disease. The scleroderma dataset has characteristics that are common in many problems where the individual unit's trajectory is the focus. We observe multivariate longitudinal measures at irregular observed times for each patient which composes the cohort. Of the continuous outcomes of interest, some follow non-Gaussian distributions. We also have multiple predictors clinically known to influence the outcomes that need to be considered. In this chapter, we focus on building models that flexibly describe the population and individual's trajectories keeping such characteristics of the data in mind.

Modeling trajectories in multiple dimensions involves a choice of whether to jointly model all markers in a single model or to model each marker in separate models. Traditionally, marker-specific models are widely used to estimate health trajectories, but the questions of whether it is advantageous to use a joint model have also been raised and answered in multiple settings over many years. In order to fully utilize information in multiple longitudinal markers, theoretically, we need to fit a joint model of all markers as opposed to marker-specific models. We present a Bayesian hierarchical model of multiple longitudinal markers to estimate population, clinical subgroups, and individuals' trajectories taking the nested structure of the data into account. The fitted model describes population disease progression in continuous time as well as correlations among markers from different organ systems.

We focus on quantifying the performance of the joint model to marker-specific models by deriving general formulae measuring the relative efficiency.

The approach is somewhat unique in that we focus on the performance of the two models on individual patient trajectories rather than the average trajectories for subpopulations. We specifically investigate inefficiencies of fitting marker-specific models for each patient and describe how they are associated with the characteristics of the patient's data. We examine in detail the implications of the formulae on the scleroderma data, but the statistical methods presented can easily be generalized for application to diverse chronic diseases where biomarkers' trajectories are of clinical importance.

2.2 Methods

2.2.1 The Johns Hopkins Scleroderma Cohort

The Johns Hopkins Scleroderma Center Cohort, one of the largest internationally, provides a unique opportunity for the development of a trajectories-based prediction tool for Scleroderma patients. In this retrospective/prospective dynamic entry cohort, the clinical states of over 4000 patients have been collected at baseline and every 6 months throughout the duration of the study that started in 1990 and continues today. The data comprises the following information:

1. demographic factors (date of birth, gender, race, ethnicity)
2. disease confirmation (classification criteria including ACR and CREST criteria) and timing of disease onset (both Raynaud's phenomenon and first non-Raynaud's symptom)
3. disease subtype (limited or diffuse based on extent of skin involvement)

4. disease severity scores (modified Medsger Severity Scale, based on 7 organ system scales)

5. exposure variables (family history, tobacco history, other exposures) and medication history

6. quality of life measures (SSc Health Assessment Questionnaire, 3 dyspnea scales, among other patient reported outcome measures)

7. history of cancer; cancer diagnosis date, site, histology, treatment (surgery, chemotherapy, radiation therapy, immunotherapy) if applicable

8. clinical and research laboratory data including autoantibody status

9. all pulmonary function tests, echocardiograms, and right heart catheterization (RHC) results.

In this paper, we model the latent health state for pulmonary function measured by the standardized percent predicted forced vital capacity (pFVC) and standardized percent predicted diffusing capacity for carbon monoxide (pDLCO), cardiac function measured by right ventricular systolic pressure (RVSP) and left ventricular ejection fraction (EF), and cutaneous involvement measured by the modified Rodnan skin score (mRSS) since individuals' disease onset. As is the clinical tradition, disease onset is defined by the earlier of the onset of Raynaud's phenomenon and first non-Raynaud's symptom.

Prior to analysis, all five measures are preprocessed using quantile normalization. Let Y_k be a vector of the observed values from each measure $k = 1, ..., 5$. The quantile-normalized vector for each k is obtained by $\hat{\Phi}^{-1} \circ \hat{G}_k(Y_k)$, where

$\hat{G}_k$ is an estimated distribution of the vector Y_k and $\hat{\Phi}^{-1}$ is the inverse of the standard normal distribution. To calculate the quantile normalized values, observations from each measure are sorted in ascending order, paired with and given the values of the corresponding percentiles calculated from a standard normal distribution.

Lastly, RVSP and mRSS are transformed by multiplying them by -1 so that increase/decrease in all five measures indicates better/worse health status. The common scale is especially useful when aggregating predicted latent health trajectories to generate a single trajectory characterizing overall health state of patients' over time.

2.2.2 Statistical models of trajectory

The linear mixed model (LMM) is widely used to describe changes in a single approximately-Gaussian longitudinal outcome over time. LMMs are commonly used in observational studies, as they describe the correlation among repeated measures of the same subjects and estimate subject-specific effects while naturally handling irregularly spaced or/and unbalanced data (e.g., Brown and Prescott, 1999; Diggle et al., 2002). In his seminal papers (Harville, 1976 and Harville, 1977), Harville established the framework of the LMM and showed that random effects estimators are the best unbiased linear predictors (BLUP) with known covariance parameters. The result is derived by extending the Gauss-Markov theorem, which was previously used to prove that the general linear mixed model (GLMM) provides the best linear unbiased estimator (BLUE) (Graybill, 1976). For a single outcome, Potthoff and Roy,

1964, Rao, 1965, and Grizzle and Allen, 1969 developed methods of estimation and inference for the regression parameters in the case of complete, balanced data. Laird and Ware, 1982 extended the LMM framework to accommodate unbalanced data captured at possibly irregular times.

The multivariate linear mixed (MLMM) is an extension of the LMM for the analysis of multiple outcomes. The MLMM was first fit using complete and balanced data in Reinsel, 1984, then Shah, Laird, and Schoenfeld, 1997 presented parameter estimation via the EM algorithm for bivariate outcomes with possibly missing responses. The variations and applications of MLMM includes Sammel, Lin, and Ryan, 1999, Fieuws and Verbeke, 2004, and Wang and Fan, 2012 among others.

In the presence of multivariate longitudinal observations measured for individuals, both LMM and MLMM are in common practice. The LMM approach estimates the population and individual trajectories of each outcome independently of the others, while the MLMM additionally captures the between-measure correlations induced by correlated random effects and random error terms. An important question is how much worse are the regression coefficients estimated from biomarker-specific LMMs performance where the correlations among the error terms are ignored relative to the corresponding coefficients estimated from a single MLMM. In early work on this question Bloomfield and Watson, 1975 derived expressions for that combinations of the design matrix and residual variance matrix for which the inefficiency of the least squares estimates compared to the BLUE is the worst possible. A similar idea was explored by Tukey, 1948. He quantified the maximum inefficiency

caused by using a misweighted mean as compared to the optimally weighted mean.

Cases in which the gain in efficiency is the largest possible have also been studied under the "seemingly unrelated regression" (SUR) framework. A SUR comprises a set of linear regression equations where each equation describes the relationship between a different outcome and its associated predictor variables. Zellner showed that the coefficient estimation using the Aitken's generalized least squares (GLS) (Aitken, 1934) is asymptotically more efficient compared to the OLS, and that the efficiency increases as the error terms from different equations become more cross-correlated and as the predictor variables in different equations become less correlated. Zellner and Huang, 1962 further established the properties of the efficient estimator and proved that the GLS also yields a minimal generalized MSE of the predicted values. Oliveira and Teixeira-Pinto, 2015 applied Zellner's result to the case in which the outcomes share some predictors in common. He showed that GLS is more efficient only for the outcome-specific predictors.

In this paper, we study the efficiency of the MLMM relative to a set of outcome-specific LMMs using the scleroderma case study as the context. We consider three parameters of interest: the fixed effects regression coefficients, the random effects for each individual, and predicted values for each interval over time. We work under the assumtion that missing data are missing at random (MAR) (Rubin, 1976). For estimation of the fixed effects parameters, we first approach the problem in the context of the SUR framework (see Section

2.2.5). We introduce a method to quantify the inefficiency of the outcome-specific LMM relative to the MLMM in the population level parameters and individual level predictions.

2.2.3 Notation

Let Y_{ijk} be the observed value for the kth measure for person $i = 1, \ldots, m$ at the jth visit $j = 1, \ldots, n_{ik}$, at time since onset t_{ijk} and, let Y_{ik} be the vector of Y_{ijk} for $j = 1, \ldots, n_{ik}$. X_{ik} and Z_{ik} are $(n_{ik} \times p_k)$ and $(n_{ik} \times q_k)$ known matrices of full rank, and β_k and b_{ik} are $p_k \times 1$ and $q_k \times 1$ measure-specific vector of parameters for fixed and random effects. Let $n_i = \sum_{k=1}^{K} n_{ik}$ and e_{ik} random measure-specific within-subject error term.

In this application, we observe K = 5 different measures:

$$k = \begin{cases} 1 & pFVC \\ 2 & pDLCO \\ 3 & EF \\ 4 & RVSP \\ 5 & mRSS. \end{cases}$$

The linear mixed effects model is written as

$$Y_i = X_i\beta + Z_i b_i + e_i,\ i = 1, ..., m$$

where

$$\beta = (\beta_1^T, ..., \beta_K^T)^T,\ Y_i = (Y_{i1}^T, ..., Y_{iK}^T)^T,\ X_i = \bigoplus_{k=1}^{K} X_{ik},\ Z_i = \bigoplus_{k=1}^{K} Z_{ik}.$$

We assume

$$b_i = (b_{i1}^T, ..., b_{iK}^T)^T \overset{ind}{\sim} N_{Kq}(0, D)$$

$$e_i = (e_{i1}^T, ..., e_{iK}^T)^T \overset{ind}{\sim} N_{n_i}(0, \Sigma_i).$$

Letting

$$Y = (Y_1^T, ..., Y_m^T)^T,\ X = (X_1^T, ..., X_m^T)^T,\ Z = \bigoplus_{i=1}^{m} Z_i$$

$$b = (b_1^T, ...b_m^T)^T,\ e = (e_1^T, ...e_m^T)^T,\ \Gamma = I_m \bigotimes D,\ \Sigma = \bigoplus_{i=1}^{m} \Sigma_i,$$

we can write the above model more compactly in the standard linear mixed model form

$$Y = X\beta + Zb + e$$

where

$$Y \sim N(X\beta, V),\ V = Z\Gamma Z^T + \Sigma$$

$$b \sim N(0, \Gamma),\ e \sim N(0, \Sigma).$$

2.2.4 Defining combined and separated models

In the above specification, D and Σ_i are $(Kq \times Kq)$ and $(n_i \times n_i)$ positive definite matrices, respectively. The K $(q \times q)$ and $(n_{ik} \times n_{ik})$ measure-specific block matrices for D and Σ_i on the diagonals represent within-measure covariance of random effects and random errors, respectively. The off block diagonals of D and Σ_i represent the covariances of random effects and random errors across measures. If they are set equal to zero matrices, then the mixed effects model of K measures reduces to what is equivalent to K measure-specific univariate mixed effects models. We call this model the "separated" model; the model with the unrestricted D and Σ_i is called the "combined" model.

For the separated model,

$$Y_i = X_i\beta_S + Z_i b_{Si} + e_i,\ i = 1, ..., m$$

where

$$b_{Si} \sim N_{Kq}(0, D_S),\ e_i \sim N_{n_i}(0, \Sigma_{Si})$$

$$Y \sim N(X\beta_S, V_S),\ V_S = Z\Gamma_S Z^T + \Sigma_S$$

$$\Gamma_S = I_m \bigotimes D_S,\ \Sigma_S = \bigoplus_{i=1}^{m} \Sigma_{Si}.$$

For the combined model,

$$Y_i = X_i\beta_C + Z_i b_{Ci} + e_i,\ i = 1, ..., m$$

where

$$b_{Ci} \sim N_{Kq}(0, D_C),\ e_i \sim N_{n_i}(0, \Sigma_{Ci})$$

$$Y \sim N(X\beta_C, V_C),\ V_C = Z\Gamma_C Z^T + \Sigma_C$$

$$\Gamma_C = I_m \bigotimes D_C,\ \Sigma_C = \bigoplus_{i=1}^{m} \Sigma_{Ci}.$$

For simpler notation, let $W_S = V_S^{-1}$, $W_C = V_C^{-1}$ and $W_{S_i} = V_{S_i}^{-1}$, $W_{C_i} = V_{C_i}^{-1}$ in following sections.

2.2.5 Separated models and seemingly unrelated regressions

The fixed effects estimates $\hat{\beta}_C$ from the combined model are generalized least squares (GLS) estimates, first described in Aitken, 1934. The Aitken model takes the form of $Y = X\beta + \epsilon$ with the first two moments of ϵ defined as $E(\epsilon|X) = 0$ and $Cov(\epsilon|X) = V$, where V is a known positive definite matrix.

Under these assumptions, the GLS estimator of β is $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$. It is the best linear unbiased estimator (BLUE); in other words, $\hat{\beta}$ is the "best" estimator in the sense that it has the minimum variance among the class of linear unbiased estimators of β. Since our model defined in 2.2.3, assumes a positive definite matrix $V = Z\Gamma Z^T + \Sigma$, we know that $\hat{\beta}_C$ is the most efficient estimators of the fixed effects assuming known variance. The result implies that even though $\hat{\beta}_S$ and $\hat{\beta}_C$ are both unbiased estimators of β as shown in Appendix B.1, there is loss of efficiency in using $\hat{\beta}_S$ as $\hat{\beta}_C$ is the most efficient.

There are, however, situations where the separated models yield estimators as efficient as those estimated in a single multivariate outcome, as introduced in Zellner, 1962. We show Zellner's results under his SUR framework and investigate whether or not the results are applicable to the estimates of fixed effects and random effects in our model.

We start with a simplified version of our model. Suppose each of our K measurements includes m individuals responses collected at times $t = 1, ...T$ and $n = m \times T$. The seemingly unrelated regression equations are defined as:

$$Y_k = X_k \beta_k + \epsilon_k,\ i = 1, ..., K$$

where Y_k and ϵ_k are $(n \times 1)$, X_k is $(n \times p)$ and β_k is $(p \times 1)$. The covariances in the errors or disturbance terms ϵ_k are defined across the five measures by $Cov(\epsilon_{kt}, \epsilon_{lt}) = \sigma_{kl} I_n$.

It is possible and often convenient to estimate the regression coefficients for each outcome separately. This is the reason that the equations are called

seemingly unrelated regression equations. However, the equations are connected to one another by the correlation among the error terms across the equations.

We can also write this model as one large linear model:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_K \end{pmatrix} = \begin{pmatrix} X_1 & 0 & 0 & \cdots & 0 \\ 0 & X_2 & 0 & \cdots & 0 \\ 0 & 0 & X_3 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & X_K \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_K \end{pmatrix} = X\beta + \epsilon.$$

With the following covariance matrix for the error term ϵ,

$$V(\epsilon) = \begin{pmatrix} \sigma_{11}I_n & \sigma_{12}I_n & \cdots & \sigma_{1K}I_n \\ \sigma_{21}I_n & \sigma_{22}I_n & \cdots & \sigma_{2K}I_n \\ \vdots & \vdots & & \vdots \\ \sigma_{K1}I_n & \sigma_{K2}I_n & \cdots & \sigma_{KK}I_n \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2K} \\ \vdots & \vdots & & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK} \end{pmatrix} \bigotimes I_n$$

$$= \Sigma \bigotimes I_n$$

σ_{kl} is the covariance between the error term in the equation for measure k and that in the equation for measure l, and σ_{kk} is the variance of the error term in the equation for measure k. The above formulation assumes that σ_{kk} and σ_{kl} are constant for all observations within respective measures, that there is no correlation between errors of observations collected at different times implying no serial correlation of error terms. More importantly, we are disregarding all within-individual correlation captured by introducing random effects. We will later accommodate our mixed effects model assumptions and modify

the results accordingly, but we first demonstrate Zellner's results using our simpler model.

The best linear unbiased estimator for β is the GLS estimates $\hat{\beta}$

$= (X^T V^{-1} X)^{-1} X^T V^{-1} Y$ where

$$V^{-1}(\epsilon) = \begin{pmatrix} \sigma^{11} I_n & \sigma^{12} I_n & \cdots & \sigma^{1K} I_n \\ \sigma^{21} I_n & \sigma^{22} I_n & \cdots & \sigma^{2K} I_n \\ \vdots & \vdots & & \vdots \\ \sigma^{K1} I_n & \sigma^{K2} I_n & \cdots & \sigma^{KK} I_n \end{pmatrix} = \begin{pmatrix} \sigma^{11} & \sigma^{12} & \cdots & \sigma^{1K} \\ \sigma^{21} & \sigma^{22} & \cdots & \sigma^{2K} \\ \vdots & \vdots & & \vdots \\ \sigma^{K1} & \sigma^{K2} & \cdots & \sigma^{KK} \end{pmatrix} \bigotimes I_n$$

$$= \Sigma^{-1} \bigotimes I_n$$

The first of the two conditions under which the equation-by-equation model yields estimators as efficient as $\hat{\beta}$ is when the error terms have a diagonal covariance matrix such that $\sigma_{kl} = \sigma_{lk} = 0$. When we force all covariance terms across measures to be 0, the GLS estimator reduces to five single-equation least-squares estimators. The less obvious condition is $X_1 = X_2 = \cdots = X_K$.

If $X_k = X_1$ for all $k = 1, \ldots, K$,

$$\hat{\beta} = \{X^T(\Sigma^{-1} \bigotimes I_n)X\}^{-1} X^T(\Sigma^{-1} \bigotimes I_n)Y = (\Sigma^{-1} \bigotimes X_1^T X_1)^{-1}(\Sigma^{-1} \bigotimes X_1^T)Y$$

$$= (\Sigma \bigotimes (X_1^T X_1)^{-1})(\Sigma^{-1} \bigotimes X_1^T)Y = (\Sigma \bigotimes (X_1^T X_1)^{-1})(\Sigma^{-1} \bigotimes X_1^T)Y$$

$$= (\Sigma\Sigma^{-1}) \bigotimes ((X_1^T X_1)^{-1} X_1^T)Y = I_m \bigotimes ((X_1^T X_1)^{-1} X_1^T)Y$$

$$= \begin{pmatrix} (X_1^T X_1)^{-1} X_1^T Y_1 \\ \vdots \\ (X_1^T X_1)^{-1} X_1^T Y_K \end{pmatrix}$$

Hence, $\hat{\beta}$ reduces to a vector of single-equation estimators even with correlated error terms. In practice, this implies that when all measures have the same design matrix (i.e. data for all K measures measured at the same time and function of time is the only explanatory variable for each measure), we obtain the same population estimates from the equation-by-equation model and the joint model. We now investigate if these two results also hold under the assumptions of the combined model.

The main difference between the SUR model and the combined model is that patient-specific random effects are introduced. The fixed effect estimates of the combined model is $\hat{\beta}_C = (X^T W_C X)^{-1} X^T W_C Y$ where $W_C^{-1} = V_C = Z\Gamma_C Z^T + \Sigma_C$. We define

$$b_{Ci} = \begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iK} \end{pmatrix} \overset{ind}{\sim} N\left(0, \begin{pmatrix} G_{11} & G_{12} & \cdots & G_{1K} \\ G_{21} & G_{22} & \cdots & G_{2K} \\ \vdots & \vdots & & \vdots \\ G_{K1} & G_{K2} & \cdots & G_{KK} \end{pmatrix}\right)$$

and

$$\Gamma_C = \begin{pmatrix} G_{11} \bigotimes I_m & G_{12} \bigotimes I_m & \cdots & G_{1K} \bigotimes I_m \\ G_{21} \bigotimes I_m & G_{22} \bigotimes I_m & \cdots & G_{2K} \bigotimes I_m \\ \vdots & \vdots & & \vdots \\ G_{K1} \bigotimes I_m & G_{K2} \bigotimes I_m & \cdots & G_{KK} \bigotimes I_m \end{pmatrix},$$

$$\Sigma_C = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2K} \\ \vdots & \vdots & & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_{KK} \end{pmatrix} \bigotimes I_n$$

Let Z_{ki} be the random effects design matrix of kth measure of ith person, $Z_k = \bigoplus_{i=1}^{m} Z_{ki}$ the design matrix for measure k, and $Z = \bigoplus_{i=1}^{K} Z_k$.

Then, V_C is

$$\begin{pmatrix} Z_1G_{11} \bigotimes I_m Z_1^T + \sigma_{11} I_n & Z_1G_{12} \bigotimes I_m Z_2^T + \sigma_{12} I_n & \cdots & Z_1G_{1K} \bigotimes I_m Z_K^T + \sigma_{1K} I_n \\ Z_2G_{21} \bigotimes I_m Z_1^T + \sigma_{21} I_n & Z_2G_{22} \bigotimes I_m Z_2^T + \sigma_{22} I_n & \cdots & Z_2G_{2K} \bigotimes I_m Z_K^T + \sigma_{2K} I_n \\ \vdots & \vdots & & \vdots \\ Z_KG_{K1} \bigotimes I_m Z_1^T + \sigma_{K1} I_n & Z_KG_{K2} \bigotimes I_m Z_2^T + \sigma_{K2} I_n & \cdots & Z_KG_{KK} \bigotimes I_m Z_K^T + \sigma_{KK} I_n \end{pmatrix}$$

The off-diagonal block matrices of V_C explaining the covariance of the observations across measures in time become 0 when we have a diagonal covariance matrix for error terms ($\sigma_{kl} = \sigma_{lk} = 0$) and additionally for random effects ($G_{kl} = G_{lk} = 0$). Then, $V_C = V_S$ and $\hat{\beta}_C = \hat{\beta}_S$, and $\hat{\beta_S}$ becomes BLUE.

The remaining question is whether $\hat{\beta}_C = \hat{\beta}_S$ in the case of $X_1 = \cdots = X_K$. As shown above, Zellner's result largely depends on the covariance of the error terms adopted in the SUR model, where it can be written as $\Sigma \bigotimes I_n$. But the form of V_C makes it clear that it cannot be decomposed into a Kronecker product of a matrix with scalar variance terms and the identity matrix. Therefore, $\hat{\beta}_C$ does not take the same form as $\hat{\beta}_S$ and $\hat{\beta}_S$ is not an efficient estimator. To conclude, estimation of the fixed effects is guaranteed to be more efficient when jointly modeling the measures as compared to separate modeling, except in the obvious case where all the covariance terms across measures are 0 in which case the two models are equivalent. Having established that separated models produce inefficient estimates of fixed effects, the question remains how inefficient are they? Is the inefficiency enough to warrant the burden of jointly modeling the outcomes in situations where the separated models meet the clinical objectives?

2.2.6 Comparing estimates of combined and separated models

Our interest lies in understanding the advantages and disadvantages of using the combined model instead of the separated model when both address the main scientific or clinical question. This is typically the case when the questions are about within-measure contrasts, rather than those that involve multiple measures. Because the combined model quantifies the covariation across measures, there are inferences about cross-measure contrasts that require use of the combined model. Examples are whether our two lung measures are sufficiently strongly associated that only one is required for clinical decision making, or to what extent disease progression in the skin forewarns future progression in the lung. In this section, we consider the relative merits with respect to accuracy and precision of the combined and separated models in estimating:

1. fixed effects coefficients that represent population average trajectories $\hat{\beta}$
2. individuals' random effects $\hat{b}_i$ that represent their deviation from the average trajectories
3. individual patients' estimated trajectories $\hat{y}_i$ that are a linear combination of $\hat{\beta}$ and $\hat{b}_i$.

Our strategy is to compare the theoretical mean squared error (MSE) and its variance and bias components for each of $\hat{\beta}$, $\hat{b}_i$, and $\hat{y}_i$ obtained from the combined and separated model. Since $\hat{\beta}$, $\hat{b}_i$, and $\hat{y}_i$ are functions of the variance estimates, we first obtain the variance estimates by fitting the combined model.

We fit a linear mixed effects model with measure-specific smooth function of time as fixed effects plus random intercept and slope. From the model, we obtain fully parametrized covariance estimates of D_C and W_{Ci}, which are assumed to be the true covariances of the underlying population. By taking block diagonal elements of D_C and W_{Ci}, we subsequently obtain D_S and W_{Si}.

The least squares estimates for fixed effects from the separated and combined models are

$$\hat{\beta}_S = (X^T W_S X)^{-1} X^T W_S Y$$

$$\hat{\beta}_C = (X^T W_C X)^{-1} X^T W_C Y.$$

The mean squared error (MSE) of an estimator $\theta \in \mathbf{R}^d$ is defined as

$$E(||\hat{\theta} - \theta||^2) = E(\sum_{j=1}^{d} (\hat{\theta}_j - \theta_j)^2) = Tr(Var(\hat{\theta})) + ||Bias(\hat{\theta})||^2.$$

Then, as both estimates are unbiased,

$$MSE(\hat{\beta}_S, \beta) = Tr(var(\hat{\beta}_S)) = Tr((X^T W_S X)^{-1} X^T W_S V_C W_S X (X^T W_S X)^{-1})$$

$$MSE(\hat{\beta}_C, \beta) = Tr(var(\hat{\beta}_C)) = Tr((X^T W_C X)^{-1}).$$

We use the conditional expectation of the random effects given the observed data for patient i as our estimates of their random effects. These conditional expectations under the combined and separated models are given by:

$$\hat{b}_{Si} = D_S Z_i^T W_{Si} (y_i - X_i \hat{\beta}_S)$$

$$\hat{b}_{Ci} = D_C Z_i^T W_{C_i} (y_i - X_i \hat{\beta}_C).$$

If we condition on the true random effect b_i, then $\hat{b}_{Si}$ and $\hat{b}_{Ci}$ are both biased

toward 0. Their MSEs are defined as the conditional expected squared difference between the predicted values above and the true value of the random effect. We calcuate the average value of the MSEs over the distribution of b_i.

For the separated model,

$$E_{b_i}\{MSE(\hat{b}_{Si},\, b_i)\} = E_{b_i}\{Tr(var_{\hat{b}_{Si}|b_i}(\hat{b}_{Si}|b_i))\} + E_{b_i}\{||Bias(\hat{b}_{Si})||^2\}$$

$$E_{b_i}\{Tr(var_{y_i|b_i}(\hat{b}_{Si}|b_i))\} = Tr\{D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si})\Sigma_{Ci}$$

$$\times (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i D_S + X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} V_{Cn} W_{Sn} X_n\}$$

$$E_{b_i}\{||Bias(\hat{b}_{Si})||^2\} = Tr\{(D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I) D_C$$

$$\times (D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I)^T\}.$$

For the combined model,

$$E_{b_i}\{MSE(\hat{b}_{Ci},\, b_i)\} = E_{b_i}\{Tr(var_{\hat{b}_{Ci}|b_i}(\hat{b}_{Ci}|b_i))\} + E_{b_i}\{||Bias(\hat{b}_{Ci})||^2\}$$

$$E_{b_i}\{Tr(var_{y_i|b_i}(\hat{b}_{Ci}|b_i))\} = Tr\{D_C Z_i^T (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci})\Sigma_{Ci}$$

$$\times(W_{Ci} - W_{Ci}X_i(X^TW_CX)^{-1}X_i^TW_{Ci})Z_iD_C + X_i(X^TW_CX)^{-1}\sum_{n\neq i}^{m} X_n^TW_{Cn}X_n\}$$

$$E_{b_i}\{||Bias(\hat{b}_{Ci})||^2\} = E_{b_i}\{Tr((E(\hat{b}_{Ci}|b_i) - b_i)(E(\hat{b}_{Ci}|b_i) - b_i)^T)\}$$

$$= Tr\{(D_CZ_i^T(W_{Ci} - W_{Ci}X_i(X^TW_CX)^{-1}X_i^TW_{Ci})Z_i - I)D_C$$

$$\times(D_CZ_i^T(W_{Ci} - W_{Ci}X_i(X^TW_CX)^{-1}X_i^TW_{Ci})Z_i - I)^T\}.$$

Lastly, predicted values for patient i under the two models are

$$\hat{y}_{Si} = X_i\hat{\beta}_S + Z_i\hat{b}_i = X_i\hat{\beta}_S + Z_iD_SZ_i^TW_{Si}(y_i - X_i\hat{\beta}_S)$$

$$\hat{y}_{Ci} = X_i\hat{\beta}_C + Z_i\hat{b}_i = X_i\hat{\beta}_C + Z_iD_CZ_i^TW_{Ci}(y_i - X_i\hat{\beta}_C).$$

MSE of the predicted values to their true trajectory for the separated model is

$$E_{b_i}\{MSE(\hat{y}_{Si}, E(\hat{y}_i|b_i))\} = E_{b_i}\{Tr(var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i))\} + E_{b_i}\{||Bias(\hat{y}_{Si})||^2\}$$

$$E_{b_i}\{Tr(var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i))\} = Tr((M_{Si}X_i^T + Z_iD_SZ_i^T)W_{Si}\Sigma_{Ci}W_{Si}(M_{Si}X_i^T + Z_iD_SZ_i^T)^T$$

$$+M_{Si}(\sum_{n\neq i}^{m} X_n^TW_{Sn}V_{Cn}W_{Sn}X_n)M_{Si}^T)$$

$$E_{b_i}\{||Bias(\hat{y}_{Si})||^2\} = Tr((\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)Z_iD_{Ci}Z_i$$

$$\times(\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)^T)$$

$$\text{where } M_{Si} = (X_i - Z_iD_SZ_i^TW_{Si}X_i)(X^TW_SX)^{-1}.$$

For the combined model,

$$E_{b_i}\{MSE(\hat{y}_{Ci},\, E(\hat{y}_i|b_i))\} = E_{b_i}\{Tr(var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i))\} + E_{b_i}\{||Bias(\hat{y}_{Ci})||^2\}$$

$$E_{b_i}\{Tr(var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i))\} = Tr((M_{Ci}X_i^T + Z_iD_CZ_i^T)W_{Ci}\Sigma_{Ci}W_{Ci}(M_{Ci}X_i^T + Z_iD_CZ_i^T)^T)$$
$$+M_{Ci}(\sum_{n\neq i}^{m} X_n^TW_{Cn}X_n)M_{Ci}^T$$

$$E_{b_i}\{||Bias(\hat{y}_{Ci})||^2\} = Tr((\{M_{Ci}X_i^T + Z_iD_CZ_i^T\}W_{Ci} - I)Z_iD_{Ci}Z_i$$
$$\times(\{M_{Ci}X_i^T + Z_iD_CZ_i^T\}W_{Ci} - I)^T)$$

$$\text{where } M_{Ci} = (X_i - Z_iD_CZ_i^TW_{Ci}X_i)(X^TW_CX)^{-1}.$$

For derivations, see Appendix A.1.

To quantify the maximum benefit of using the combined model, we first calculated the combined/separated ratio of measure-specific mean squared errors, biases, and variances when the variance parameters are assumed to be known. We used the estimated posterior modes for the unknown variance parameters as if they were the true values. We then accounted for the fact that the variance parameters are unknown and must be estimated. It is expected that some or all of the advantage of the combined model might be lost. We approximated the posterior distribution of the MSEs, biases, and variances. For each of the 500 random samples from the joint posterior, we calculate the errors, biases, and variances of the two models using the formulae above. We then compare the ratio posterior distributions to the previously obtained ratios that ignored the uncertainty in the variance parameters.

2.2.7 Efficiency gained by the degree of correlation across patient-specific trends

Our conjecture is that the measures that are heavily correlated are the ones that benefit the most from fitting the combined model and measures that have weaker correlation will only have marginal gains. We further test this idea by gradually amplifying the correlation of one measure's random effect to those of other measures and investigate the effect of higher correlation on efficiency gain for that measure.

We vary the degree of correlation among random effects for the EF and rest of the measures as follows. From the combined model, we obtain a correlation matrix of random effects:

$$C_b = (diag(D_C))^{-\frac{1}{2}} D_C (diag(D_C))^{-\frac{1}{2}}.$$

In C_b, we increase the absolute value of the correlation between the EF random slope and those of all other measures to 0.1, 0.2, and 0.3 while retaining the sign of the original correlation. To ensure the resulting covariance matrices are positive-semi-definite, we use a slightly modified version of the spectral decomposition method introduced in Rebonato and Jäckel, 2001. Let's call the modified correlation matrices $C_{0.1}$, $C_{0.2}$, and $C_{0.3}$. We decompose each matrix into its eigenvalues and eigenvectors. For each set of eigenvalues, we replace any negative values by half of the smallest positive eigenvalue to obtain $C^*_{0.1}$, $C^*_{0.2}$, and $C^*_{0.3}$. Then, the resulting covariance matrices $\hat{D}_{0.1}$, $\hat{D}_{0.2}$, $\hat{D}_{0.3}$ are calculated as

$$\hat{D}_C = (diag(D_C))^{\frac{1}{2}} C^*_b (diag(D_C))^{\frac{1}{2}}$$

where C_b^* is $C_{0.1}^*$, $C_{0.2}^*$, and $C_{0.3}^*$ respectively. The correlations between EF random slope and the 4 other measures in $C_{0.1}^*$, $C_{0.2}^*$, and $C_{0.3}^*$ might not have the magnitude of exactly 0.1, 0.2, and 0.3, respectively and the calculated correlations are shown in the results section and appendix.

2.2.8 Fitting the separated and combined models

In order to investigate the advantages and disadvantages of the two models as they might be used in statistical practice, we sought to fit the models using existing open-source software. An appropriate software package should accommodate the following conditions: (1) the number of repeated measurements and the times of measurement can differ across the subjects and the measures without requiring explicit imputation of outcomes and (2) the covariance across random effects and measurements can be flexibly constructed to represent the correlations observed in the scleroderma case-study data.

To reflect changes in patients' health over time since disease onset, the fixed effects of our model included natural splines of time with 3 degrees of freedom, age of onset, race, sex, skin type, presence of three common autoantibodies, and the interactions of each of the baseline covariates listed above with the natural spline of time. Patient specific intercept and linear time are included as random effects. Standard linear mixed model software including R packages **lme4** (Bates et al., 2015) and **nlme** (Pinheiro et al., 2019) can easily fit the separated models. However, in this case-study, the algorithm fails to converge despite substantial efforts to tailor the convergence tuning constants. The combined model with saturated random effects and residual

covariances requires estimation of 40 + 10 additional parameters in the random effects and residual covariance matrices, respectively, compared to those of the separated model.

A second alternative was to use a Bayesian Markov Chain Monte Carlo (MCMC) algorithm, where we can obtain posterior distributions of all parameters of interest. We successfully fit the combined model using an R package **MCMCglmm** (Hadfield, 2010). For the fixed effects of both models, we use a diffuse independent normal prior centered around zero with a large variance (10^8). Weakly informative inverse-Wishart priors are placed on random effects and residual covariance matrices. After examining the distribution of our data, we set the prior distribution of the random intercepts to have the mode of 1 and those of random slopes to have the mode of 0.005, with 10 degrees of freedom. The prior distribution of the residual covariance matrix takes the mode of 1 for each measure, with 5 degrees of freedom. The degrees of freedom are chosen to make the distributions as diffuse as possible while guaranteeing them to be valid inverse-Wishart distributions. Scale matrices for both distributions are calculated using the chosen degrees of freedom and modal values.

2.3 Results

2.3.1 Data description

For the following analyses, we include 581 "data-rich" patients who have at least 4 data points from disease onset to 40 years later for each of the 5 measurements. Table 2.1 presents the mean and standard deviation of the

number of observations per patient.

	pFVC	pDLCO	EF	RVSP	mRSS
Mean	12.83	12.40	9.13	7.47	19.09
StdDev	6.20	6.01	3.71	3.28	7.61
n	6136	5789	4281	3281	9055

Table 2.1: Summary statistics of patients' number of observations

Table 2.2 summarizes the distribution of the continuous and binary variables that are included as covariates in the model.

	(N = 581)
Age of Onset	
minimum	3.60
median (IQR)	41.65 (31.61, 51.17)
mean (sd)	41.42 $\pm$ 14.02
maximum	83.45
Race	
African American	91 (15.7%)
Other	490 (84.3%)
Type	
Diffuse	254 (43.7%)
Other	327 (56.3%)
Sex	
Female	497 (85.5%)
Male	84 (14.5%)
ACA	
-	450 (77.5%)
+	131 (22.5%)
RNA pol	
-	500 (86.1%)
+	81 (13.9%)
Scl-70	
-	449 (77.3%)
+	132 (22.7%)

Table 2.2: Characteristics of the cohort

2.3.2 Estimating latent trajectory

2.3.2.1 Population average trajectory

From the simulated parameter values from the posterior distribution, we calculate population mean trajectories for each of the five measures. Figure 2.1 (bottom) presents the posterior mean biomarker value as a function of time with 95% credible intervals defined to be the interval from the 2.5 to 97.5 percentiles of the posterior draws. We observe a small decline in pFVC, a large decline pDLCO, a small decline in the heart measure RVSP, but a strong improving trend for the skin condition mRSS.

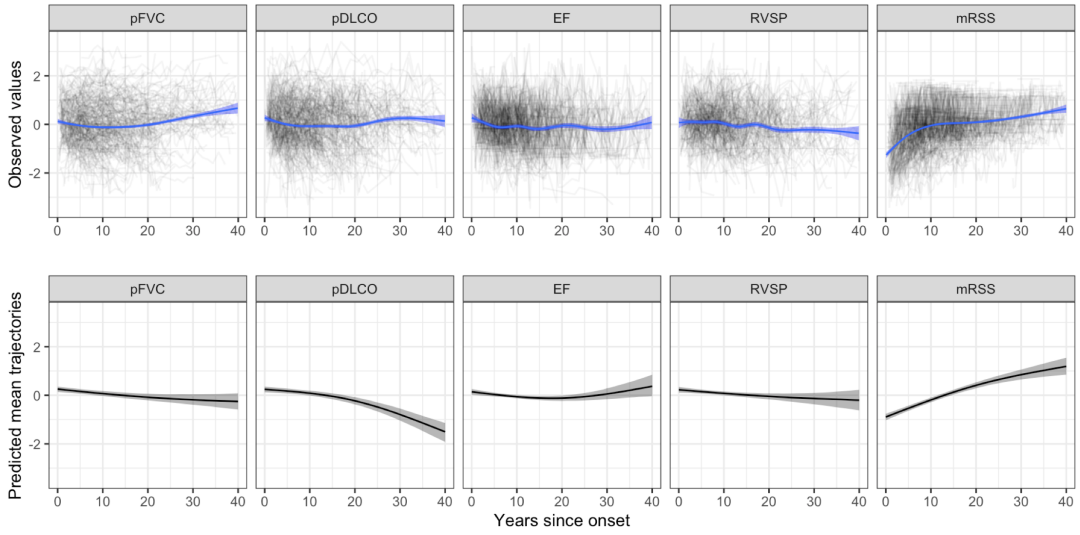


Figure 2.1: Observed and predicted population average trajectories

The set of five plots in the top row of Figure 2.1 shows the preprocessed observed data. Each individual's longitudinal measurements are shown as grey lines in the background. The blue smoothing lines connect the empirical mean value at each time point with nominal 95% confidence intervals that

incorrectly ignore the correlations among the repeated measurements for an individual. We observe upward trend in the observed population trajectory for pFVC and pDLCO while the predicted trajectories of the two measures show decreasing trend. This can be explained by selection bias; patients with more negative trajectories drop out more often. The observed trends only reflect the average of observed values of the remaining healthier patients.

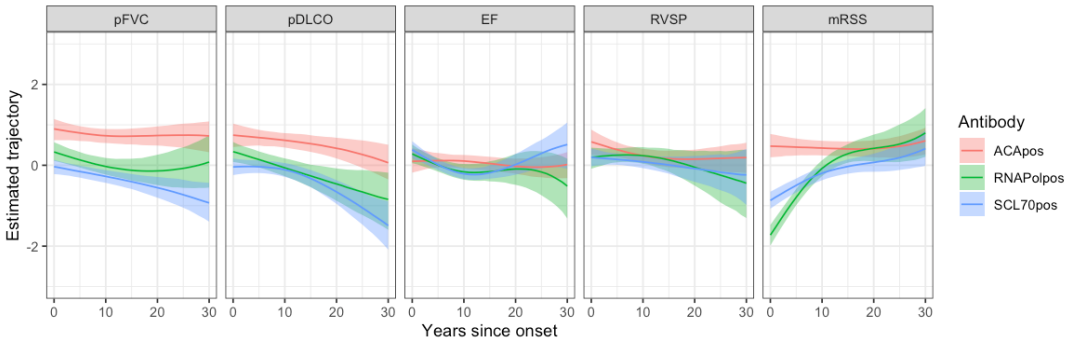


Figure 2.2: Predicted average trajectories for autoantibody groups

The health trajectories serve as a useful tool to estimate disease prognosis for scientifically-defined subgroups. We estimate average health trajectories for three autoantibody groups by fitting the combined model. In Figure 2.2, the predicted mean trajectories and their 95% credible intervals for the five measures are shown for each patient subgroup: those positive for anti-centromere antibodies (ACA), anti-RNA polymerase (RNAPol), and anti-Scl-70 antibodies (Scl-70). The autoantibody subgroups demonstrate distinct paths most notably in the two lung measures pFVC and pDLCO and in the first 10 years since onset for mRSS. Patients tested positive for ACA tend to have better prognosis for lung and skin. RNAPol and Scl-70 groups have lower initial levels and steeper negative slopes for the lung measures and positive slopes

for mRSS - indicating more advanced skin and lung disease at disease onset with progressive lung disease over time and improving cutaneous disease.

2.3.3 Separated and combined models

2.3.3.1 Measure-wise correlation

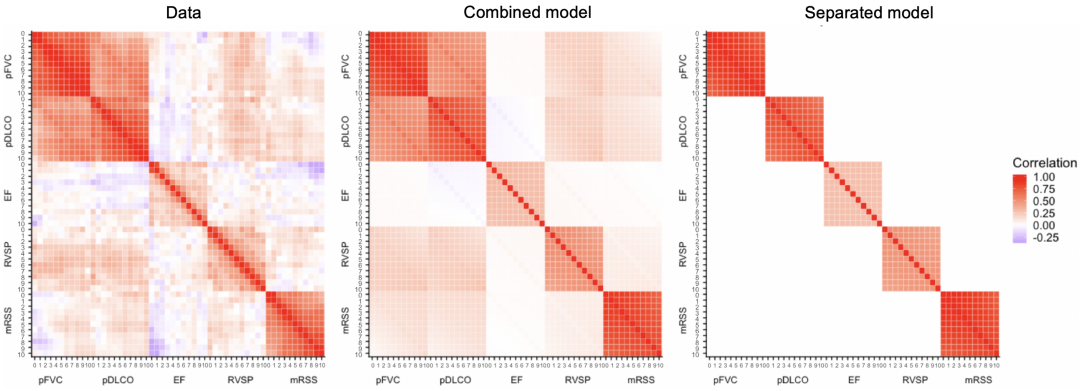


Figure 2.3: Empirical correlation matrix

We display the correlation across the observed measures at different times in the data and compare the correlation structures assumed and estimated by the two models. Pairwise correlations of observations from all patients for 11 years (years 0,...,10 since the disease onset) are calculated and plotted (Figure 2.3 (left)) using range of colors from red, white, and blue each representing correlation of 1, 0, and -1, respectively. The 11 by 11 block matrices on the diagonals shows the degree of correlation in patients' repeated observations over time for each of the five measures. Looking along the block-diagonal, one observes that the two lung measurements and mRSS are highly correlated with their respective past observations, while observations of the two heart measures have less serial correlation. The 11 by 11 off-diagonal block matrices

show the degree of correlation for different pairs of measures with leads and lags. We observe high positive correlation for the two lung measures which suggests that there could be gains in efficiency when modeling the two measures jointly. We observe some degree of positive correlation between RVSP, mRSS, and the two lung measures; the EF observations appear to be uncorrelated with any other measure including the other cardiac measure.

The estimated random effects correlation matrix C_b' calculated from $\hat{D}_C$ captures the associations among random levels and linear trends for each pair of measures. In order to investigate the degree of correlation of patient-specific trends, we present Table 2.3, which only includes the random slope components of C_b^*. We observe a correlation of 0.64 between the latent trends of pFVC and pDLCO. The correlations of RVSP random slope with that of pFVC and pDLCO are estimated to be 0.37 and 0.40, respectively. mRSS random slope also have positive correlation, 0.27 and 0.20, with pFVC and pDLCO random slope. The EF trend has little or no correlation with pFVC and RVSP trends and small correlations with pDLCO and mRSS.

In Figure 2.3 (middle and right), the empirical correlation matrices of the combined and separated models are plotted using the covariance estimates from the two models, $\hat{D}_C$ and $\hat{D}_S$. Recall that the combined model allows correlation among the five measures, while the separated model does not. As shown in Figure 2.3, the combined model captures patterns in the within-measure and across-measure empirical correlation matrices of the data quite well, while the separated model only captures within-measure correlations.

	pFVC	pDLCO	EF	RVSP	mRSS
pFVC	1.00	0.64	0.00	0.37	0.27
pDLCO	0.64	1.00	-0.12	0.40	0.20
EF	0.00	-0.12	1.00	-0.03	0.10
RVSP	0.37	0.40	-0.03	1.00	0.13
mRSS	0.27	0.20	0.10	0.13	1.00

Table 2.3: Correlation of random slopes across measures from C_b

2.3.3.2 Comparing bias and efficiency

Using the formulas derived in 2.2.6 we compare MSE, bias and variance components of: (1) the fixed effects estimates $\hat{\beta}_C$ and $\hat{\beta}_S$; (2) random effects estimates $\hat{b}_{Ci}$ and $\hat{b}_{Si}$; and (3) the predicted values $\hat{y}_{Ci}$ and $\hat{y}_{Si}$. All three estimands of interest are functions of the design matrices (X and Z) and covariance matrices (D_C, D_S, Σ_{Si}, and Σ_{C_i}). We construct design matrices for each individual using observed times at which the five measurements are taken based on the model described in 2.2.8. From the model, we also estimate the population covariance of the random effects D_C and population residual covariance Σ_{Ci}. In this section, we use the finite sample posterior estimates of the theoretical variances obtained by taking the posterior mean of the MCMC estimates of D_C and Σ_{Ci}. D_S and Σ_{Si} are constructed by forcing the off-diagonal terms of D_C and Σ_{Ci} to be 0, indicating the absence of across-measure variance terms of the separated model.

Population average trajectory estimation

Efficiency gain in estimating population health trajectory can be evaluated by comparing the variance components of fixed effects estimates of the two models. In Table 2.4, we present $MSE(\hat{\beta}_C, \beta)/MSE(\hat{\beta}_S, \beta)$, the ratio of MSE of

overall and measure-wise fixed effects of the combined model to the separated model.

	Overall	pFVC	pDLCO	EF	RVSP	mRSS
MSE Ratio$_\beta$	0.97	0.98	0.97	0.99	0.95	0.99

Table 2.4: Ratio of MSE of overall and measure-wise fixed effects of the combined model to the separated model

Assuming known variances, the overall MSE in estimating fixed effects is reduced by only 3% when using the combined model compared to fitting the separated model. Since both fixed effect estimates for the separated and combined models ($\hat{\beta}_S$ and $\hat{\beta}_C$) are unbiased (see Appendix B.1), the reduction in MSE solely comes from a variance reduction. The comparative advantage is more apparent for RVSP but not for EF. Clinically, the result implies that we need 3% more data when fitting the separated model compared in estimating population mean trajectories as efficiently as the combined model, and 5% more data for RVSP alone. This small gain in efficiency does not account for the uncertainty in the variance parameters so is not likely to be worth the increased modeling demands of fitting a combined model.

Estimating random effects

Patients' deviations in the level and trend from the average population trajectory is captured by the random intercept and slope estimates. As shown in Section 2.2.6, the random effects estimates are a linear combination of the patient-specific level and trajectory estimates and the population estimates. Hence, depending on the amount and characteristics of individual's data, we can expect variation among patients in the MSE, squared bias, and variance.

In this section, we first present the overall effect of fitting the combined model on the random effects and random slope estimates by calculating the average of patients' three estimates for each of the five measures.

The ratios of the three estimands are calculated as the following:

$$\text{MSE Ratio}_{b_i} = E_{b_i}\{MSE(\hat{b}_{Ci}, b_i)\}/E_{b_i}\{MSE(\hat{b}_{Si}, b_i)\}$$

$$\text{Squared Bias Ratio}_{b_i} = E_{b_i}\{||Bias(\hat{b}_{Ci})||^2\}/E_{b_i}\{||Bias(\hat{b}_{Si})||^2\}$$

$$\text{Variance Ratio}_{b_i} = E_{b_i}\{Tr(var_{\hat{b}_{Ci}|b_i}(\hat{b}_{Ci}|b_i))\}/E_{b_i}\{Tr(var_{\hat{b}_{Si}|b_i}(\hat{b}_{Ci}|b_i))\}$$

	pFVC	pDLCO	EF	RVSP	mRSS
Mean MSE Ratio$_{b_i}$	0.98	0.96	0.98	0.95	0.99
Mean Squared Bias Ratio$_{b_i}$	1.00	0.97	0.98	0.94	0.99
Mean Variance Ratio$_{b_i}$	0.99	1.01	1.01	1.06	1.00

Table 2.5: Mean and the ratio of MSE, variance, and squared bias components of random effects of the combined model to the separated model

In Table 2.5, we observe modest reductions in the MSE for the combined model. We also observe that the advantage is mostly generated from reduction in squared bias rather than variance. The reduced squared bias suggests that the random effects estimators from the combined model are closer to the true random effects on average compared to those from the separated model. Similar to the population average trajectory estimation, we observe greater average decrease in MSE and Squared Bias for RVSP than for the other measures.

We also compare the means of three estimands only for the random slopes of the two models in Table 2.6, where we observe similar patterns. Estimating

random slopes using the combined model is most advantageous for RVSP with smaller average MSE and squared bias.

	pFVC	pDLCO	EF	RVSP	mRSS
Mean MSE Ratio$_{b_i}$	0.97	0.95	0.98	0.91	0.98
Mean Squared Bias Ratio$_{b_i}$	1.00	0.97	0.97	0.91	0.99
Mean Variance Ratio$_{b_i}$	0.99	1.01	1.03	1.08	1.00

Table 2.6: Mean and the ratio of MSE, variance, and squared bias components of random slopes (trends) of the combined model to the separated model

Individual patients' prediction

Individual's predicted trajectory is a combination of estimated population mean trajectories and individual specific trajectory fitted only through the individual's data points. The amount of shrinkage to the group mean from an individual's observed trajectory depends on the amount of her data. At the limit of no observations, the best predicted trajectory for an individual is the population mean trajectory, itself.

The ratios of the three estimands are calculated as follows:

$$\text{MSE Ratio}_{y_i} = E_{b_i}\{MSE(\hat{y}_{Ci},\, E(\hat{y}_i|b_i))\} / E_{b_i}\{MSE(\hat{y}_{Si},\, E(\hat{y}_i|b_i))\}$$

$$\text{Squared Bias Ratio}_{y_i} = E_{b_i}\{||Bias(\hat{y}_{Ci})||^2\} / E_{b_i}\{||Bias(\hat{y}_{Si})||^2\}$$

$$\text{Variance Ratio}_{y_i} = E_{b_i}\{Tr(var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i))\} / E_{b_i}\{Tr(var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i))\}$$

As was the case for the fixed effects, the mean gains in MSE by fitting the combined model are minimal as shown in Table 2.7. The mean Squared Bias Ratio and mean Variance Ratio are also qualitatively not different than was observed for the fixed effects. Of course, the degree of benefit obtained

from the combined model will vary among individuals. The size of this heterogeneity is considered next.

	pFVC	pDLCO	EF	RVSP	mRSS
Mean MSE Ratio$_{y_i}$	0.99	0.97	0.99	0.97	0.99
Mean Squared Bias Ratio$_{y_i}$	1.02	1.01	0.99	1.02	1.00
Mean Variance Ratio$_{y_i}$	0.99	0.97	0.99	0.96	0.99

Table 2.7: Mean of ratio of MSE, variance, and squared bias components of predicted values of the combined model to the separated model

2.3.3.3 Heterogeneity in bias and efficiency gains by patient

In this section, we focus on the values and distribution of the MSE, bias and variance for estimators of patients' means, random effects and predicted values. The MSE and its components are calculated using (1) asymptotic variance estimates and (2) the posterior distribution of the variance estimates. All patient-specific expected errors are summarized in Figure 2.4. Boxplots for errors using the asymptotic variance are plotted in grey, and those using the posterior distribution in orange. Since the MSE, bias, and variance are transformed onto the log scale, a positive value indicates that the separated model has smaller errors and a negative value indicates that the combined model does.

As using the asymptotic variances yields a single population level estimate for each measure, the point estimates are plotted in grey points in the top three plots. The five grey points marking the measure-specific log ratios of MSE (left) and variance (right) are equivalent to the MSE Ratio$_\beta$ in Table 2.4 transformed to the log scale. The box plots in orange show the distribution of the 500 estimands by each measure calculated using the 500 posterior MCMC

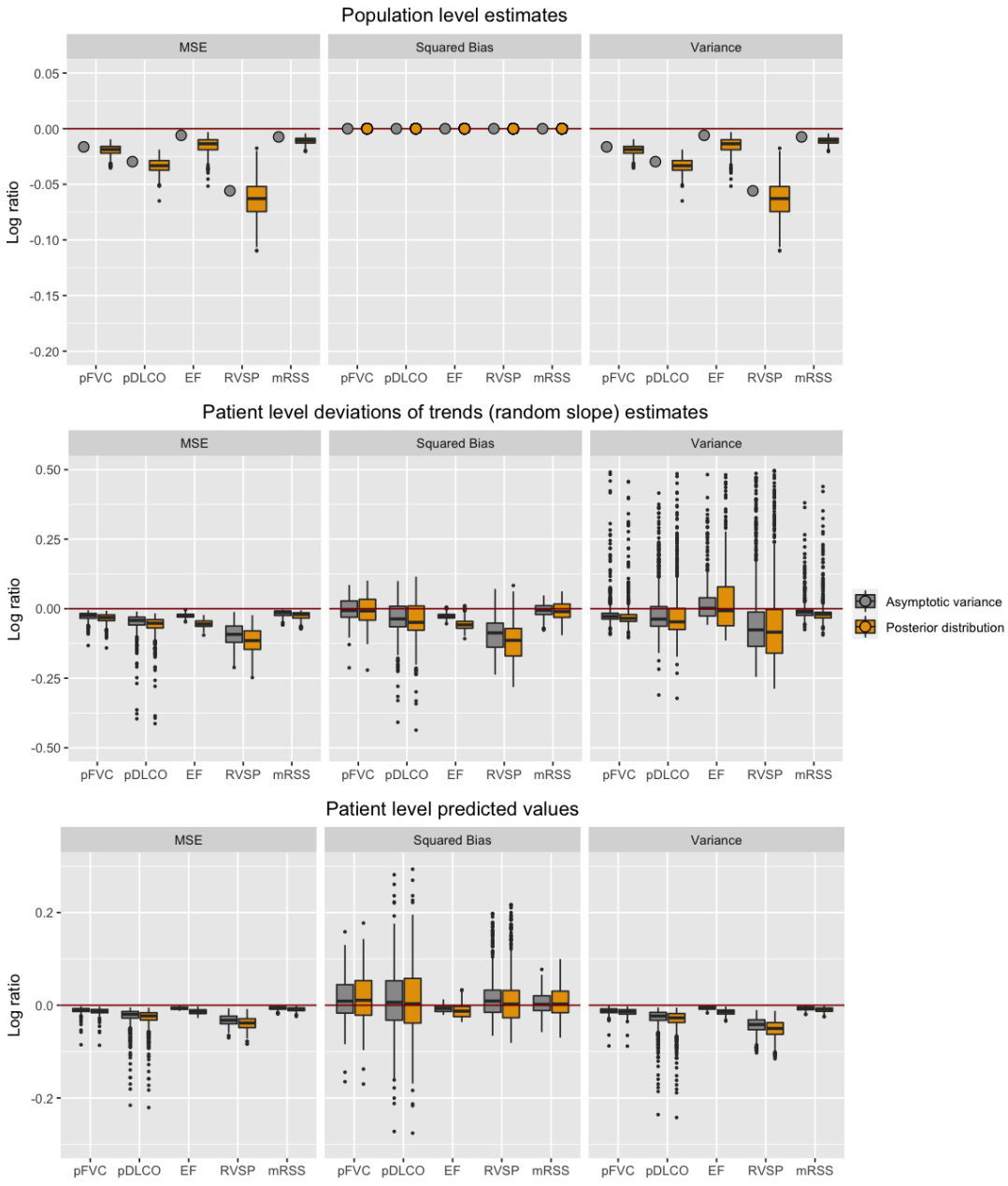


Figure 2.4: Log ratio of MSE, variance and squared bias components of the combined model to the separated model

samples for the variance estimates. We observe that the grey point estimates are generally close to the center of the orange box plots. As the squared bias

components are 0 for both measures, the log ratios are plotted at 0.

We also compared the patient-level log ratios for the random slope estimates and predicted values. Three plots in the middle and bottom row show measure-specific log ratios of 581 patients. The grey box plots characterize the distributions of the log ratios using asymptotic variances. When using the posterior variance samples, each patient has a distribution of the 500 copies of the three estimands. To summarize the information in these distributions, we take the mean of the 500 ratios and plot the logged values shown across patients in orange box plots.

The most notable result is that there is a sizable amount of heterogeneity for the patients' log ratios, especially for pDLCO and RVSP. Most patients benefit from the reduction in RVSP MSE ratios by fitting the combined model. The gains in pDLCO are more noticeable for the 25th percentile of the patients. The stretched out left tails of the pDLCO MSE for both random slope estimates and predicted values indicate that some patients are estimated to have over 20% ($\approx e^{-0.2}$) efficiency gains. The factors that influences the varying degrees of efficiency gains across patients are further investigated in the next section.

Overall, the reduction in MSE results from the reduction in squared bias for the random slope estimates and from the reduction in variance for the predicted values. Comparing the magnitude of log ratios, the combined model reduces the MSE the most for the random slope estimates compared to the population estimates and predicted values. We conclude that there hardly is a qualitative difference in the results from using asymptotic variance estimates and using the posterior distribution of the variance estimates.

2.3.3.4 Gains in efficiency by patient data characteristics

In this section, we present a series of post hoc analyses that investigate the association between efficiency gains and patient data characteristics. From previous sections, we observed that the reduction in MSE at all levels - population and individual - is the greatest for RVSP. We can intuitively relate this result to the number of observations in each measure. In Table 2.1, RVSP has the lowest average number of observations per patient. On average, the two lung measures and mRSS have twice as much data as in RVSP. Hence, utilizing information in the richer measures by fitting the combined model is likely to result in improved estimation of the RVSP parameters.

Richness of data, however, cannot solely explain the efficiency gains. Both EF and RVSP have sparse data, yet EF hardly benefits from fitting the combined model at any level. Unlike RVSP, the degree of correlation between EF and other measures are close to 0 as shown in Table 2.3. Using the method described in Section 2.2.7, we increase the magnitude of correlation between EF random slope and those of all other measures to 0.1, 0.2, and 0.3 as described in Table 2.8. The estimated correlation matrices across random slopes with varying degrees of correlations are in Appendix A.2.

	pFVC	pDLCO	RVSP	mRSS
Corr Data	0.00	-0.12	-0.03	0.10
Corr0.1	0.10	-0.10	0.10	-0.10
Corr0.2	0.20	-0.20	0.20	-0.20
Corr0.3	0.28	-0.28	0.29	-0.28

Table 2.8: Correlation of EF random slope and other measures' random slopes from C_b, $C'_{0.1}$, $C'_{0.2}$, and $C'_{0.3}$.

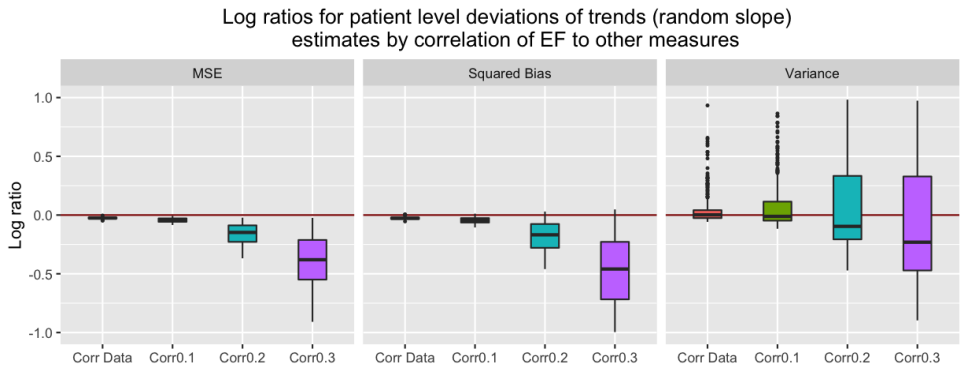


Figure 2.5: Log ratio of MSE, Variance, and Squared Bias of patient level deviations of trends of the combined to separated model by varying degrees of correlation in random slope of EF and other measures.

The effects of varying correlations on the EF log ratios are presented in Figure 2.5. As the correlations of EF random slope and those of other measures increase from the unmodified correlation estimated from the data (Corr Data) to 0.3, we observe considerable reductions in MSE and squared bias. The result illustrates that fitting the combined model is especially advantageous when the patient level trends for one measure are more strongly correlated with the others.

Finally, we present the influence of information in one measure on estimation of the other in the presence of across-measure correlation. We focus on pDLCO and pFVC, whose random slopes are highly correlated at 0.64 (Table 2.3). In Figure 2.6, we illustrate pDLCO patient level log ratios by the number of pDLCO and pFVC observations. Generally, the MSE ratios for the random slope estimates and predicted values are lower for patients with relatively fewer pDLCO observations. The result implies that patients with rich pDLCO data can obtain reasonable trajectory estimates by only modeling pDLCO,

while patients with sparse pDLCO data are likely to benefit from improved estimation by fitting the combined model.

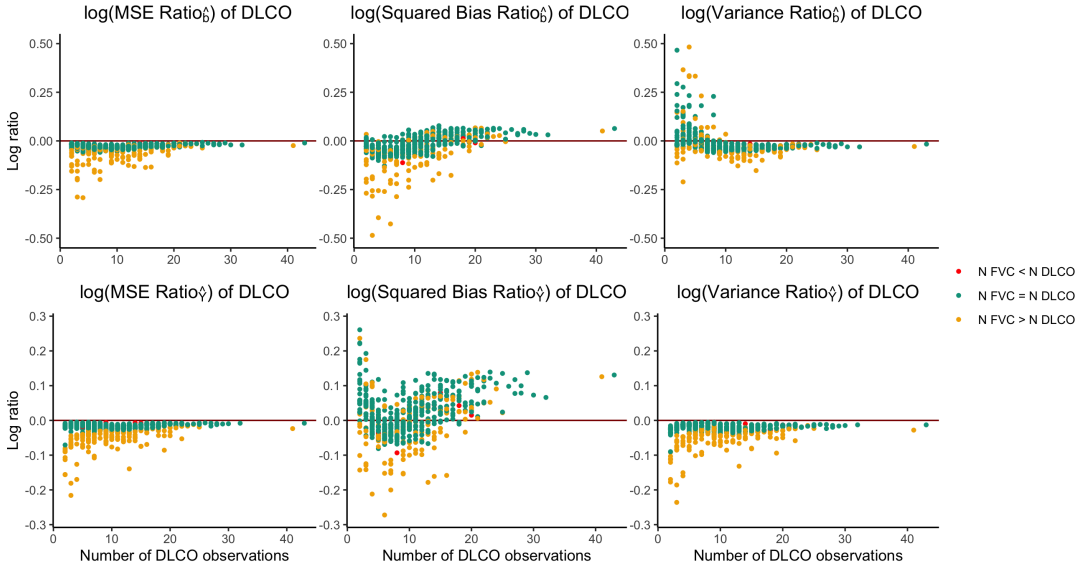


Figure 2.6: Patient level log ratios of MSE, variance and squared bias components of the combined model to the separated model for pDLCO.

Among these patients with fewer pDLCO observations, we identify the subset of patients who benefit the most from fitting the combined model by stratifying them into three groups by the relative number of pFVC and pDLCO observations. We notice lower log MSE ratios for patients whose number of pFVC observations is greater than that of pDLCO (points plotted in yellow in Figure 2.6), whereas there are only minimal gains for those whose number of pFVC observations is equal to or lower than that of pDLCO (points plotted in green and red). From these results, we conclude that the available information in the measure itself and other correlated measures together determine how much an individual can benefit from fitting the combined model.

2.4 Discussion

In this paper, we present our approach to estimating latent health trajectory from multivariate longitudinal data. We define and compare the combined and separated models with the aim of providing guidelines for modeling when individuals' trajectory is reflected in multiple longitudinal outcomes. By the Gauss-Markov theorem, the combined model is known to yield more efficient fixed and random effects estimates. However, the combined model is also more complex and potentially less robust to misspecification. Our main question is to what degree the assumed benefits outweigh the simplicity of relying on separated models. We first extended the SUR framework to the MLMM and showed that the separated model can only be as efficient as the combined model if there is no correlation between the random effects and random disturbance terms across different measures. Then we provided an approach to quantify the efficiency gain in estimating the population level parameters as well as the individual level predictors. We give general formulae for the relative efficiency of estimators for the fixed effects, random effects and predicted values. In our case study, we observe minimal gains in efficiency for the fixed effect estimates by fitting the combined model. Gains in estimating the random effects was the largest when the given measure is highly correlated with other measures, and when the relative number of the observations in that measure is smaller than those in other correlated measures.

In such cases, we observe gains in efficiency, that is smaller MSE, that mostly results from reduced bias. For individuals who have only a few data points available for a given measure, the data for the measure alone cannot

accurately reflect the underlying disease state of the individual, and fitting the separated model results in more shrinkage towards the measure-specific mean and hence larger bias. The bias is reduced when fitting the combined model, where the random effects estimator borrows strength from data-rich measures. We also show that the results hold under the assumption of known variance parameter and when we take the uncertainty in the variance into account. Although we present the inefficiency in estimation that comes from the misspecification of variance in the attempt to answer specific medical questions pertaining to scleroderma, the approach can be flexibly generalized to other applications. We proposed a framework to compare the performances of the combined and separated models for the population and individual level estimates, which can be applied to any setting where the individuals' and population trajectory in higher dimension space need to be estimated. However, it should be noted that the results are drawn assuming normal population with MAR assumption. The effect on disparities from normality and non-ignorable missingness on the results remains to be studied.

Chapter 3

Predicting clinical events using Bayesian multivariate linear mixed models

3.1 Introduction

It is a major challenge to assess risks of critical events in chronic, multi-organ diseases such as multiple sclerosis (Institute of Medicine, 2001), lupus (Zeller and Appenzeller, 2008), and Parkinson's disease (Jain, 2011). Scleroderma, an autoimmune disease that is manifested by fibrosis of multiple organ systems, may affect the skin, heart, lungs, kidneys, gastrointestinal tract, and blood vessels. Severe organ involvement can result in early death (Pattanaik, Brown, and Postlethwaite, 2011; Steen and Medsger, 2000). The 9-year cumulative survival rate for diffuse scleroderma patients with severe organ involvement was estimated to be 38% (Steen and Medsger, 2000). Mortality is highest due to pulmonary and cardiac complications of the disease; 35% of scleroderma-related death has been attributed to pulmonary fibrosis, 26% to pulmonary arterial hypertension (PAH) and 26% to cardiac causes (Tyndall Anthony J. et

al., 2010). Such events are commonly observed in scleroderma patients; for example, pulmonary involvement has been reported in up to 25% of patients at the early stage of diagnosis (Mcnearney et al., 2007). Hence, a major goal is the early detection of patients who are most likely to progress at an early stage of the disease, as this may provide a window of opportunity to intervene before there is irreversible organ damage (Shah and Wigley, 2013).

In monitoring scleroderma, clinicians obtain longitudinal markers from pulmonary function tests and echocardiograms to assess whether there is evidence of disease activity or progression in the lung and heart. Left ventricular ejection fraction (EF), right ventricular systolic pressure (RVSP), and percent predicted forced vital capacity (pFVC) are examples of parameters that are monitored to detect whether there is emerging cardiomyopathy, pulmonary hypertension (PH) and ILD, respectively. For each of these measures, a value above or below clinically established thresholds is a surrogate for these endpoints. The situation for managing scleroderma patients is common to many other chronic diseases. Multiple measurements are observed on each organ system, discrete events need to be identified, and measurements are highly irregular in their distributions and observation times. By accurately estimating an individual's organ-specific trends using such data, we can explain how the disease is evolving over time and also provide probabilities of the patient having one or more critical events in the near future. Additionally, we can potentially add value to patient care by fitting models that jointly estimate disease trajectories for several organs. It should be noted that, in this study,

the events of interest are solely determined by the values of continuous measurements. In other words, the clinical events of interest are the biomarkers themselves crossing a threshold. For cases where the events are not direct functionals of continuous observations (for example, death or renal crisis of scleroderma patients), jointly modeling longitudinal and time-to-event data to predict the risk of having an event given longitudinal profiles are widely used. The joint model proposed by Faucett and Thomas, 1996 and Wulfsohn and Tsiatis, 1997 are considered as the standard models, and several extensions were proposed to accommodate multivariate longitudinal profiles. Xu and Zeger, 2001 used a multivariate mixed model framework to model multiple continuous surrogate markers to evaluate treatment effect in a schizophrenia trial and proposed a measure to quantify the relative benefits of using multiple surrogates. Rizopoulos and Ghosh, 2011 proposed a semiparametric multivariate joint model to model 3 longitudinal outcomes time to renal graft failure. Other applications are presented by Brown, Ibrahim, and DeGruttola, 2005, Proust-Lima and Taylor, 2009, and Garre et al., 2008.

In the following sections, we introduce and fit multivariate models to estimate individual's risk of the critical events that serve as surrogates of cardiomyopathy, PH, and ILD. Similar to the models mentioned above, we will jointly fit a model with multiple longitudinal outcomes of interest. We expect better performance of this joint model as the model will borrow strength from the observations in other markers in predicting risk in less frequently measured markers. We compare the precision of our multivariate model to a series of models that use the events themselves as the outcome and takes the

individuals' past trajectory as predictor variables. We further investigate the gains in precision by fitting our multivariate model in different situations, i.e. where outcome measures are highly correlated compared to the case where they are not. Finally, we present our approach to checking the violation of the normality assumptions of the multivariate models and tests to check the need of model calibration.

3.2 Methods

3.2.1 Modeling multivariate measures and events

Motivated by the scleroderma case study, we propose to develop methodology to infer individual and population etiologies by fitting Bayesian multivariate hierarchical models introduced in Chapter 2 that describe the relationship between an individual's disease trajectory as reflected in the combination of sentinel events and longitudinal measures.

In Figure 3.1, η_i represents the time-varying disease state of patient i that is reflected in a vector of observations Y_i and events E_i. The underlying health state η_i are measured with measurement error ϵ_i whose covariance is $\Sigma_{\eta,Y}$. At any given time t, the observed covariates including baseline measurements X_i such as autoantibody status and cutaneous subtypes cause the disease trajectory status η_i as quantified by regression coefficients $\beta_{\eta,X}$. Population-level estimates $\beta_{\eta,X}$ combined with an individual's random deviation from the population average b_i fully describes the individual's trajectory at any given time.

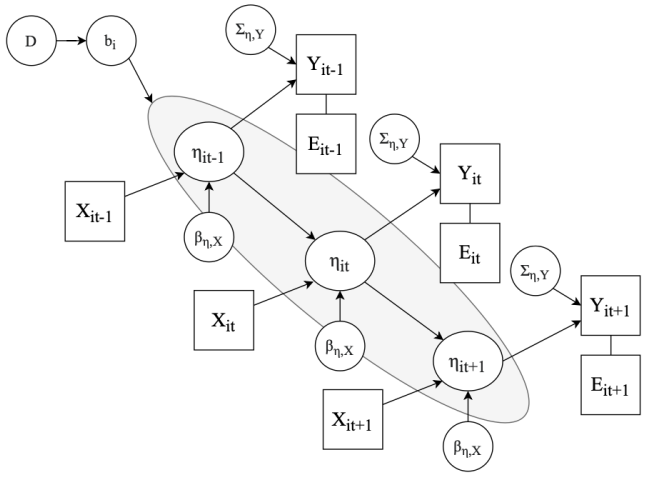


Figure 3.1: Diagram describing etiology of disease progression of an organ reflected in a single biomarker

The model can capture the disease state in multiple organs by jointly fitting all biomarkers in a single model at once. Univariate analyses in which each outcome measure or event is considered on its own are more popular largely because modeling a single outcome, even several times, is much easier than modeling them jointly. However, in such models, the across-measure associations in the random effects and residual errors are ignored. Failure to account for these associations results in less efficient estimators particularly in the presence of high measure-wise correlations as shown in Chapter 2.

3.2.2 Multivariate outcome models

In this section, we define the longitudinal measurements and clinical events used for this study. We use clinical data in the Johns Hopkins Scleroderma Center Research Registry. Longitudinal data of the biomarkers ejection fraction (EF), right ventricular systolic pressure (RVSP), percent predicted forced vital

capacity (FVC), and percent predicted diffusing capacity of carbon monoxide (DLCO) are used to describe the disease trajectory of patients who have at least 3 observations for each of the measures. Onset of the disease is defined as the earlier of the onset of Raynaud's phenomenon and first non-Raynaud's symptom. We restrict our analysis to data collected between 0 to 40 years since onset. Clinicians define thresholds for EF, RVSP, and FVC events below which the patient is said to experience: cardiomyopathy, pulmonary hypertension (PH), and interstitial lung disease (ILD), respectively. We use the following two thresholds for each measure to differentiate between mild and severe events:

$$E_{EF} = I\{EF < 50\} \text{ and } I\{EF < 35\}$$

$$E_{RVSP} = I\{RVSP \geq 45\} \text{ and } I\{RVSP \geq 50\}$$

$$E_{pFVC} = I\{pFVC \leq 70\} \text{ and } I\{pFVC \leq 60\}$$

From the empirical correlation matrices in Figure 3.2, we observe that RVSP observations are generally highly correlated with DLCO and FVC, while EF is not. We compare the performance of the predictors $p(E_{pFVC})$ from two models to assess benefit gained by jointly modeling more highly correlated variables. We fit two multivariate linear mixed models each with three longitudinal outcomes. The first model uses pFVC, pDLCO, and EF, and the second model uses pFVC, pDLCO, and RVSP. For any patient at a given moment in the future, $p(E_{EF})$ and $p(E_{FVC})$ can be calculated from the first model, and $p(E_{RVSP})$ and $p(E_{FVC})$ from the second.

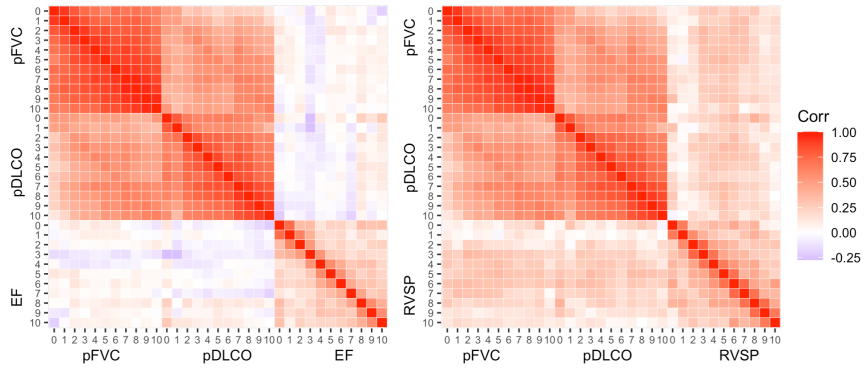


Figure 3.2: Empirical correlation matrix of the preprocessed variables

For each outcome measure, we select a set of predictors (covariates) for the fixed effects. The common predictors across all outcomes are age of scleroderma onset, race, gender, skin subtype, and autoantibody status for the 3 most common scleroderma specificities (ACA, RNAPol and Scl-70). To model changes in patients' health trajectories in time since onset, we also include a smooth function of time using natural splines with 3 degrees of freedom where internal knots are placed at 10 and 30 years since onset, and boundary knots at 0 and 40 years since onset. Note that the set of common variables are allowed to have different coefficients for each measure. Measure-specific regression predictors and coefficients are essential to describe the state of a patient's scleroderma, as it is known that each clinical subtype is at different risks for organ complications (Shah and Wigley, 2013). For example, patients with limited skin type are at higher risk of developing PH but lower risk of developing ILD (Schoenfeld and Castelino, 2015; Legendre and Mouthon, 2014).

For patient-specific random effects, we fit a random slope and intercept and two linear splines at 3 and 10 years from the last observation. The same

set of variables are used as random effects for all outcome variables. Covariate estimates for the two spline terms represent additional rate of change in a patient's trajectory in the last 10 years and then last 3 years. These terms are introduced to capture the recent trend in the trajectory more accurately, by not letting the observations measured early in the disease have excessive influence on the recent trend.

3.2.3 Preprocessing of longitudinal data

Prior to analysis, all 4 outcome measures are preprocessed using quantile normalization. Let Y_k be a vector of the observed values from each measure $k = 1, ..., 4$. The quantile normalized vector is obtained by $\Phi^{-1} \circ \hat{G}_k(Y_k)$, where $\hat{G}_k$ is an estimated distribution of the vector Y_k and Φ^{-1} is the inverse of the standard normal distribution. To calculate the quantile normalized values, observations from each measure are sorted in ascending order, paired with and then assigned the values of the corresponding percentiles from standard normal distribution. Lastly, RVSP observations are transformed by multiplying by -1 so that increase in all 4 measures indicates better health status.

Thresholds for the three events are also transformed to the normalized scale, which we will call c_{EF}, c_{RVSP}, and c_{pFVC}. Also, note that each measure is individually transformed to follow a standard normal distribution, but this procedure does not guarantee joint normality of the random errors and random effects of the 3 variables used in each model. In section 3.2.7, we propose a simple method to check whether the joint normality assumption is

seriously violated.

3.2.4 The multilevel response models and prediction

Let Y_{ijk} be the observed value for the kth measure for person $i = 1, \ldots, m$ at the jth visit $j = 1, \ldots, n_{ik}$, at time since onset t_{ijk} and, let Y_{ik} be the vector of Y_{ijk} for $j = 1, \ldots, n_{ik}$. X_{ik} and Z_{ik} are $(n_{ik} \times p_k)$ and $(n_{ik} \times q_k)$ known matrices of full rank, and β_k and b_{ik} are $p_k \times 1$ and $q_k \times 1$ measure-specific vector of parameters for fixed and random effects. Let $n_i = \sum_{k=1}^{K} n_{ik}$ and e_{ik} random measure-specific within-subject error term.

In this application, we observe K = 3 with five different measures.

$$\text{In model 1, } k = \begin{cases} 1 & pFVC \\ 2 & pDLCO \\ 3 & EF \end{cases} \quad \text{and in model 2, k} = \begin{cases} 1 & pFVC \\ 2 & pDLCO \\ 3 & RVSP \end{cases}$$

Each linear mixed effects model is written as

$$Y_i = X_i\beta + Z_i b_i + e_i,\ i = 1, ..., m$$

where

$$\beta = (\beta_1^T, ..., \beta_K^T)^T,\ Y_i = (Y_{i1}^T, ..., Y_{iK}^T)^T,\ X_i = \bigoplus_{k=1}^{K} X_{ik},\ Z_i = \bigoplus_{k=1}^{K} Z_{ik}$$

and we assume

$$b_i = (b_{i1}^T, ..., b_{iK}^T)^T \overset{ind}{\sim} N_{Kq}(0, D)$$

$$e_i = (e_{i1}^T, ..., e_{iK}^T)^T \overset{ind}{\sim} N_{n_i}(0, \Sigma_i).$$

Let Y_{ij+} be the $K \times 1$ vector of patient i's health state at an unobserved time t_{ij+}. Future observations Y_{ij+} satisfy the model assumption

$$Y_{ij+} = X_{ij+}\beta + Z_{ij+}b_i,\ i = 1, ..., m$$

where X_{ij_+k} and Z_{ij_+k} are $(1 \times p_k)$ and $(1 \times q_k)$, and $X_{ij+} = \bigoplus_{k=1}^{K} X_{ij_+k}$, $Z_{ij+} = \bigoplus_{k=1}^{K} Z_{ij_+k}$. The random errors $e_{ij+} \overset{ind}{\sim} N_K(0, \Sigma_{ij+})$. To predict the probability of clinical events at t_{ijk}, we use the conditional distribution of Y_{i+} given Y_i. Here, Y_i is the vector of outcomes of patient i observed until t_{ijk}. The joint distribution of Y_i and Y_{ij+} is

$$\begin{pmatrix} Y_i \\ Y_{ij+} \end{pmatrix} \sim N\left(\begin{pmatrix} X_i\beta \\ X_{ij+}\beta \end{pmatrix}, \begin{pmatrix} V_i & C_{ij+} \\ C_{ij+}^T & V_{ij+} \end{pmatrix}\right)$$

where $V_i = Z_i D Z_i^T + \Sigma_i, V_{ij+} = Z_{ij+} D Z_{ij+}^T + \Sigma_{ij+}$, and $C_{ij+} = Z_i D Z_{ij+}^T$

Hence, $Y_{ij+}|Y_i = y_i \sim N\left(E(Y_{ij+}|Y_i = y_i),\ Var(Y_{ij+}|Y_i = y_i)\right)$ and

$$E(Y_{ij+}|Y_i = y_i) = X_{ij+}\beta + C_{ij+}^T V_i^{-1}(y_i - X_i\beta)$$

$$Var(Y_{ij+}|Y_i = y_i) = V_{ij+} - C_{ij+}^T V_i^{-1} C_{ij+}$$

When no outcome is observed before t_{ijk} , $E(Y_{ij+}|Y_i = y_i)$ and $Var(Y_{ij+}|Y_i = y_i)$ reduce to $E(Y_{ij+}) = X_{ij+}\beta$ and $Var(Y_{ij+}) = V_{ij+}$. We can obtain a set of probabilities at t_{ijk}, by calculating the probability of the conditional distributions falling below c_{EF}, c_{RVSP}, and c_{pFVC}.

From model 1,

$$P(E_{EF,ij_+}) = \Phi(c_{EF}, E(Y_{ij_+3}|Y_i = y_i), Var(Y_{ij_+3}|Y_i = y_i))$$

$$P(E_{pFVC,ij_+}) = \Phi(c_{pFVC}, E(Y_{ij_+1}|Y_i = y_i), Var(Y_{ij_+1}|Y_i = y_i))$$

From model 2,

$$P(E_{RVSP,ij_+}) = \Phi(c_{RVSP}, E(Y_{ij_+3}|Y_i = y_i), Var(Y_{ij_+3}|Y_i = y_i))$$

$$P(E_{pFVC,ij_+}) = \Phi(c_{pFVC}, E(Y_{ij_+1}|Y_i = y_i), Var(Y_{ij_+1}|Y_i = y_i))$$

where $\Phi(x, \mu, \sigma^2)$ is a normal density function with mean μ and variance σ^2.

3.2.5 Bayesian inference

The probabilities given in the previous section are just functionals on the parameters of the multivariate mixed effects model. We estimate the posterior distribution of the model parameters and of our functionals of interest using a Bayesian inference framework. We fit models 1 and 2 using an R package **MCMCglmm** (Hadfield, 2010) and obtain posterior distributions of all parameters of interest from Markov Chain Monte Carlo (MCMC) chains. For the fixed effects of both models, we use a diffuse independent normal prior centred around zero with a large variance of 10^8. Weakly informative inverse-Wishart priors are placed on random effects and residual covariance matrices. Based upon prior medical knowledge, we set the prior distribution of the random intercepts to have mode equal to one and random slopes and linear spline terms to have the mode equal to 0.005, with 12 degrees of freedom. The

prior distribution of the residual covariance matrix takes the mode of 1 for each measure, with 3 degrees of freedom. The degrees of freedom are chosen to make the distributions as diffuse as possible while guaranteeing them to be valid inverse-Wishart distributions. Scale matrices for both distributions are calculated using the chosen degrees of freedom and modal values.

3.2.6 Cross-validated sequential prediction (CVSP) for multivariate longitudinal data (MLD)

The model described above yields the predicted probability of having an event for any unobserved outcome measured at t_{ijk}. We want to compare to our predictions with the observed rates of events. For each patient i, we could naively use all observed data Y_i until t_{ijk} and untruncated data of all other patients but doing so requires fitting the models $m \times n_{ik} \times K$ since observation times for the different measures are irregular. Refitting the models whenever a new data point is collected is extremely inefficient especially in a clinical setting.

One way to circumvent this problem is to calculate $E(Y_{ij+}|Y_i = y_i)$ and $Var(Y_{ij+}|Y_i = y_i)$ using $\hat{D}$ and $\hat{\Sigma}_i$ estimated from a pool of patients' excluding patient i. Once posterior means for $\hat{D}$ and $\hat{\Sigma}_i$ are obtained, $\hat{V}_i$, $\hat{V}_{ij+}$, $\hat{C}_{ij+}$, and $\hat{\beta}$ are easily calculated as they are functionals of D and Σ_i. As we sequentially move from t_{i1k} to $t_{in_{ik}k}$, we generate time-varying X_{ij+} and Z_{ij+} and obtain the predicted event probabilities for all n_i time points with pre-estimated $\hat{D}$ and $\hat{\Sigma}_i$. This approach, which we will call Cross-validated Sequential Prediction (CVSP for MLD) is also used to assess model precision. We perform 5-fold cross-validation by dividing all patients into 5 groups. 5 sets of $\hat{D}$ and $\hat{\Sigma}_i$

are estimated using 80% of the data not in the selected fold, and $\hat{P}(E_{EF,ij+})$, $\hat{P}(E_{RVSP,ij+})$, and two $\hat{P}(E_{pFVC,ij+})$ are calculated for all $\sum_i n_i$ measurements for the patients in the selected fold. Finally, we calculate and compare the cross-validated AUC (CV-AUC) of the 4 estimated event probabilities.

3.2.7 Checking of joint normality assumption

Although the outcome variables are quantile normalized individually, there is no guarantee that the normalized outcomes follow a multivariate normal distribution. As our prediction models depend on the joint normality assumption of the random effects and random errors, we need to check whether this assumption is reasonable, and when it is not, have a method to calibrate the predictions to better match the cross-validated observed rates of events. Here, we first introduce a method of checking systematic or extreme departures that may affect the performances of prediction by examining the marginal residuals from models 1 and 2 described in 3.2.2.

Recall the notation

$$Y = (Y_1^T, ..., Y_m^T)^T,\ X = (X_1^T, ..., X_m^T)^T,\ Z = \bigoplus_{i=1}^{m} Z_i$$

$$b = (b_1^T, ...b_m^T)^T,\ e = (e_1^T, ...e_m^T)^T,\ \Gamma = I_m \bigotimes D,\ \Sigma = \bigoplus_{i=1}^{m} \Sigma_i$$

so that our joint model has the simple form

$$Y = X\beta + Zb + e$$

$$\text{where } Y \sim N(X\beta, V),\ V = Z\Gamma Z^T + \Sigma,\ b \sim N(0, \Gamma),\ e \sim N(0, \Sigma).$$

With known β, the residuals from the linear regression models $Y - X\beta = Zb + e$ which is normal with mean 0 and covariance matrix V. We examine the normality of scaled residuals of each measure by calculating $U = diag(\hat{V})^{-1/2}(Y - X\hat{\beta})$ which should be samples of standardized normal. We examine the Q-Q plots for each measure where the standardized residuals are plotted against the standard normal and look for obvious departures of the points from the 45 degree line.

3.2.8 Empirical prediction models

In order to compare the performance of our proposed approach implemented using CVSP, we propose alternative, simpler prediction methods using logistic regressions. We build a set of models to predict EF events and then another set for RVSP. The three models LM1, LM2, and LM3 are defined as follows:

$$LM1 : logit(E_{ij+}) = ns(Y_{prev1}, \nu) + Y_{ij,pFVC} + Y_{ij,pDLCO} + \text{common covariates}$$

$$LM2 : logit(E_{ij+}) = ns(Y_{prev1}, \nu) + ns(Y_{prev2}, \nu) + Y_{ij,pFVC} + Y_{ij,pDLCO}$$

$$+ \text{common covariates}$$

$$LM3 : logit(E_{ij+}) = \sum_{j<j+} E_{ij} + ns(Y_{prev1}, \nu) + ns(Y_{prev2}, \nu) + Y_{ij,pFVC} + Y_{ij,pDLCO}$$

$$+ \text{common covariates.}$$

Here, $logit(E_{ij+})$ is the logarithm of the odds of having an EF or RVSP event at time $j+$ for patient i. As we are not directly modeling the latent trajectory as we did in the two models described in Section 3.2.4, we sequentially add

covariates that can summarize past trajectories in multiple measures. Y_{prev1} and Y_{prev2} are the most recent and second to the most recent observations of EF or RVSP prior to $j+$ for patient i. We fit a smooth function of Y_{prev1} and Y_{prev2} using natural splines with $\nu = 2$ degrees of freedom. $Y_{ij,pFVC}$ and $Y_{ij,pDLCO}$ are patient i's most recent observations of pFVC and pDLCO prior to $j+$. $\sum_{j<j+} E_{ij}$ is the counts of EF or RVSP events in the past for patient i before $j+$. We expect the additional information of past trajectory reflected in Y_{prev2} and $\sum_{j<j+} E_{ij}$ to result in improved prediction. The common covariates are identical to those in Section 3.2.4. For each of the two outcomes, we calculate CV-AUC from 5-fold cross-validation of the three models.

3.2.9 Calibration of CVSP for MLD

We check if calibration is needed for our models by examining predicted probabilities of 6 events. For each set of predicted probabilities, we compare the estimated and observed cases within quintiles of the predicted probabilities. We perform chi-square goodness of fit tests to test whether the estimated number of cases matches the cases we observe across the quintiles. If we fail to reject the null, indicating a discrepancy between the estimated and observed probabilities, then we calibrate the model using the logistic regression of the observed cases against a smooth function of predicted values.

3.3 Results

Table 3.1 shows the number of patients whose data are used in model 1 and 2 and the number of events occurred. There are 577 patients with more than

3 observations for EF, pFVC, and pDLCO and 459 patients with more than 3 observations for RVSP, pFVC, and pDLCO. EF events are rarer than RVSP events; FVC events are the most common.

Model 1	n	$EF < 50$	$EF < 35$	$pFVC \leq 70$	$pFVC \leq 60$
	577	173	36	2215	1183
Model 2	n	$RVSP \geq 45$	$RVSP \geq 50$	$pFVC \leq 70$	$pFVC \leq 60$
	459	381	240	1891	995

Table 3.1: Number of patients and events used in model 1 and model 2

We first performed a series of Chi-square goodness of fit tests using the observed and expected cases in the quintiles of predicted probabilities for each event from all proposed models. Tables 3.2 and 3.3 present the Chi-square statistics and p-values. For all models and events, we failed to reject the null hypothesis that there is statistically significant difference between the observed counts and expected counts of events estimated from the models. Hence, we proceed without calibration of the predicted probabilities and compare the precision of the proposed models.

	CVSP	LM1	LM2	LM3
$EF < 50$	6.74 (p = 0.76)	13.51 (p = 0.98)	13.41 (p = 0.98)	6.31 (p = 0.72)
$EF < 35$	9.49 (p = 0.91)	31.44 (p = 1.00)	18.52 (p = 1.00)	10.51 (p = 0.94)
$RVSP \geq 45$	6.33 (p = 0.72)	7.5 (p = 0.81)	7.8 (p = 0.83)	1.57 (p = 0.09)
$RVSP \geq 50$	9.05 (p = 0.89)	8.71 (p = 0.88)	11.36 (p = 0.96)	3.17 (p = 0.33)

Table 3.2: Results from Chi-square goodness of fit test of the proposed models

	CVSP from model 1	CVSP from model 2
$FVC \leq 70$	97.16 (p = 1.00)	120.89 (p = 1.00)
$FVC \leq 60$	13.34 (p = 0.98)	11.55 (p = 0.96)

Table 3.3: Results from Chi-square goodness of fit test of CVSP from model 1 and 2

Table 3.4 presents the CV-AUC for the two EF events and two RVSP predicted by the CVSP and three empirical methods. For the three empirical methods, we observe that sequentially adding covariates that summarizes individuals' past history results in improved or similar (comparing AUC for LM1 and LM2 in predicting $EF < 35$) prediction. The CVSP yields the highest CV-AUC in predicting all events. Comparing within the EF and RVSP events, the CVSP demonstrates better performance for stricter thresholds ($EF < 35$ and $RVSP \geq 50$) whereas the three empirical methods show similar precision. The result suggests that the CVSP may be useful in predicting other rare clinical events.

	CVSP	LM1	LM2	LM3
$EF < 50$	0.822 (0.785-0.858)	0.762 (0.660-0.864)	0.779 (0.682-0.876)	0.788 (0.693-0.862)
$EF < 35$	0.830 (0.752-0.909)	0.772 (0.729-0.816)	0.771 (0.728-0.815)	0.794 (0.755-0.836)
$RVSP \geq 45$	0.855 (0.835-0.875)	0.790 (0.764-0.817)	0.798 (0.772-0.824)	0.811 (0.786-0.837)
$RVSP \geq 50$	0.872 (0.848-0.895)	0.796 (0.762-0.829)	0.802 (0.769-0.835)	0.817 (0.785-0.848)

Table 3.4: Cross-validated AUC and 95% CI of 4 critical events by proposed methods

There is only minimal difference of the two models in predicting pFVC events using CVSP. For $pFVC \leq 70$, CV-AUC is 0.956 (95% CI: 0.951-0.961) for model 1 and 0.955 (95% CI: 0.950-0.960) for model 2. For $pFVC \leq 60$, CV-AUC is 0.961 (95% CI: 0.956-0.967) for model 1 and 0.960 (95% CI: 0.954-0.966) for model 2. The two models performing equally well implies that the estimated pFVC trajectory coupled with that of pDLCO are highly predictive of the pFVC events. The trend is captured in Figure 3.2, where we can observe highly correlated pFVC measurements across time within individuals unlike EF or RVSP. It is likely that jointly modeling pFVC and pDLCO leaving out the cardiac measurements can also produce a highly predictive model.

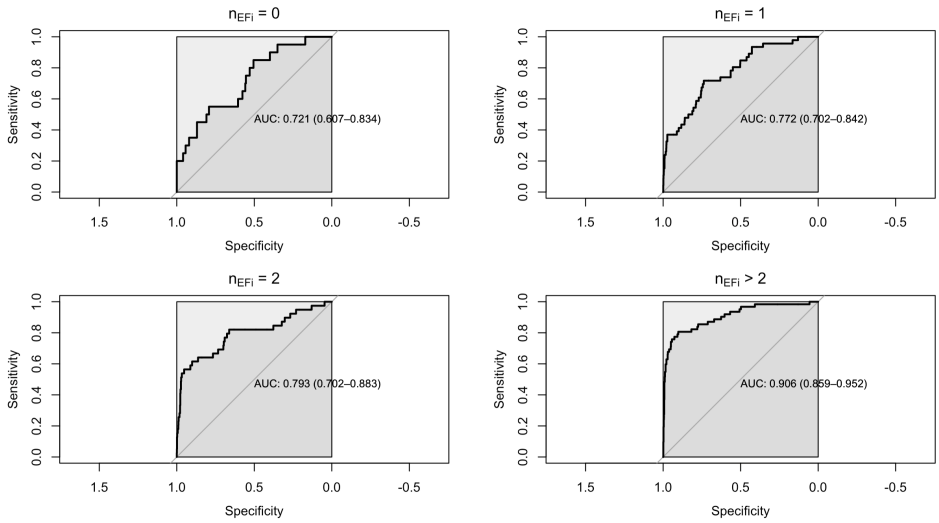


Figure 3.3: Cross-validated AUC by the number of EF observations. n_{EFi} indicates the number of EF measurements observed for individual i prior to making a prediction.

Figure 3.3 shows how much the CVSP for MLD improves over time as more data are observed. The CV-AUC improves from 0.721 to 0.906 as the number of the observed EF measurements increase from 0 to over 2, suggesting that the more data a patient has, the better precision is expected. However, even in the case of no previous observations, the CVSP has decent precision, illustrating that patients' demographic and clincal subtype along with their estimated pFVC and pDLCO trajectories provide reasonable prediction of their future EF trajectory.

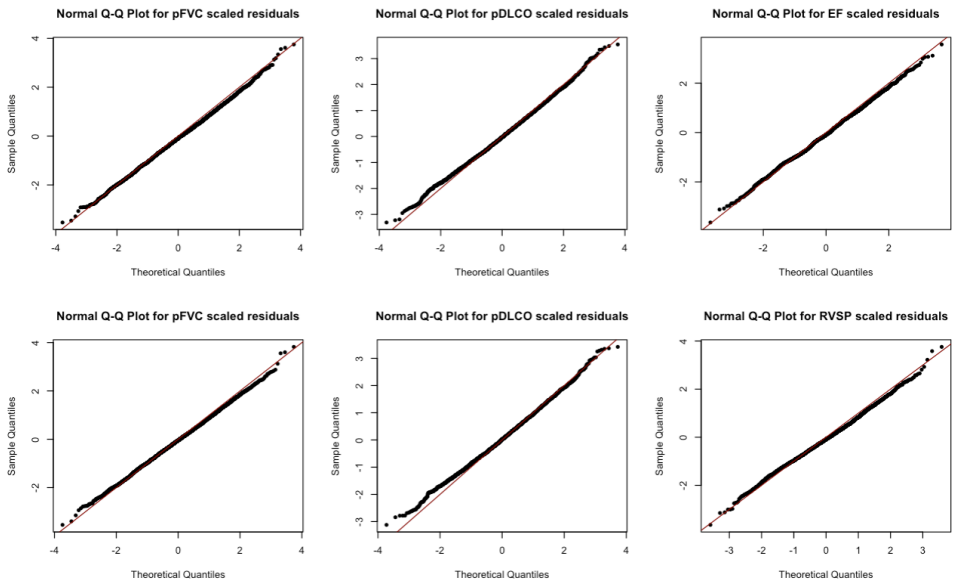


Figure 3.4: Normal Q-Q Plots

Finally, Figure 3.4 shows the Q-Q plots of the standardized residuals against the standard normal. The three plots on the top row show the sample quantiles of each of the measures in model 1 against quantiles of standard normal, and the three plots on the bottom row show those from model 2. We conclude that there are no major departures in the distributions of the scaled residuals from normality that can significantly compromise the CVSP predictions as the results above confirm.

3.4 Discussion

We introduced Bayesian multivariate hierarchical models to quantify an individual's risk of multiple important clinical events. The events are defined by clinicians to be the crossing of a biomarker threshold. So there is potential advantage to modeling the multivariate biomarkers themselves, and then

use the model to predict the crossings. We demonstrated that this approach does produce predictions with substantially higher precision as compared to the traditional prediction models. We also showed that the proposed models better separate patients when there is longer follow-up time, but we have considerable precision even with shorter follow-up time. In this analysis, we focused on the marginal risk of individual events, but we can easily obtain the joint predicted probabilities of multiple events and estimates of other quantities of interest from the models' joint posterior distributions.

For scleroderma patients, cardiomyopathy, PH, and ILD are events with high morbidity and mortality, and timely risk predictions are essential because they: (1) warn clinicians of higher risk in need of increased monitoring and interventions; (2) reduce concerns in patients at lower risk. The method of multivariate sequential updating of predictive distributions has broad application in the clinical setting. The method can be easily implemented as it provides individualized latent disease trajectories and risks of future events in multiple organs as patients' new data are observed without requiring refitting the prediction models. To use CVSP for MLD, we divided patients into K = 5 folds of data. To obtain predictions for a patient in the kth fold, we produce estimates of the random error and random effects covariance matrices leaving out the kth fold. In this way the estimated matrices by construction are independent of the left out data. We then use the fixed and random effect estimates which are functionals of the estimated covariance matrices to sequentially estimate a patient's risk of events given only the past data. Although we demonstrated our approach with an application to scleroderma,

it has broader application to other complex diseases that require multiple measures to monitor progression.

We check the normality assumptions of random effects and random error terms to see if the assumptions are seriously violated by investigating the scaled residuals. For our particular data, we did not observe any major departure in the distributions of the residuals from normality. We also confirmed that there was no obvious need to perform calibrations of the prediction models from a series of chi-square goodness of fit tests. However, in order for our models to be clinically used, it is important to thorough rigorous variable selection and model validation. We have selected the common covariates that are known to influence these outcome variables based on clinical knowledge, but the model allows independent fixed and random effects for each outcome variable including time-varying covariates. There are other modeling choices such as more informative prior distributions or alternative covariance structure specification for random errors and random effects that also needs to be investigated to improve our proposed models.

Chapter 4

Patient interface design and evaluation

4.1 Background and Significance

Systemic sclerosis is a complex autoimmune rheumatic condition that can involve multiple organ systems (skin, peripheral vasculature, heart, lung, kidneys, muscles, joints, etc). It is standard of care to assess these multiple organ systems separately – for instance, by capturing measures of skin thickness, distribution and extent (the modified Rodnan skin score or mRSS), Raynaud's phenomenon severity (digital pitting scars, ulcerations, and gangrene), left and right ventricular function and estimated pulmonary arterial pressures from echocardiograms, and forced vital capacity and diffusing capacity measurements from pulmonary function tests. Medical decision making is quite challenging because it requires integrating information across multiple parameters and organ systems, factoring in a patient's prior trajectory and baseline risk factors, and deciding whether a therapeutic intervention is warranted and if so, what is the optimal treatment. Aggregating this complex, longitudinal

data for clinical use prior to clinic visits requires a tremendous time investment on the part of the treating physician. It is also challenging to clearly explain this information to patients during a routine clinical visit to facilitate shared decision making. Lastly, because scleroderma is a complex and rare disease, it is often difficult to address questions of importance to patients, such as: what is the current status of my disease; what is my future likely to hold; and how do I compare with other scleroderma patients?

Modern statistical methods discussed in prior chapters can use multivariate, longitudinal measures to address patient's questions. To do so, we must design, implement and test an interactive interface to communicate visualizations of a patient's and reference population's longitudinal data. The first step is to construct a web-based application that estimates and communicates a person's likely disease status, past trajectory, and predictions of what is expected in the coming period. Then, to improve clinical practice, it must be tested in our clinics, and finally implemented within the workflow of the health systems that use it.

In this study, we present the design of our data visualization tools and propose a clinical trial to test their efficacy. Specifically, the trial will address two questions:

- In comparison to the distribution of states and trajectories for a reference scleroderma population, does visualization of a patient's individual disease state and trajectory improve patient satisfaction and comfort with medical decision making, and reduce decisional regret?
- Does access to a patient level data visualization model improve the

efficiency of care, in terms of data capture, delivery of information to patients, and physician satisfaction?

Once our visualization tools are implemented at Johns Hopkins and approved for wider use based upon the clinical trial designed here, we will enhance the tool by including predictions of future trajectories and the risks of organ-specific complications as described in the previous chapter. We will use a similar protocol to test the utility of the enhancements. We will then seek to disseminate this tool beyond Johns Hopkins for broader use in the rheumatology community.

4.2 Methods

4.2.1 The Johns Hopkins Scleroderma Center Research Registry

This study will use clinical data in the Johns Hopkins Scleroderma Center Research Registry, that is a university supported Precision Medicine Center of Excellence (PMCOE). The Center includes all patients who meet classification criteria for systemic sclerosis (SSc, scleroderma) who are seen at the Johns Hopkins Scleroderma Center or are inpatients at Johns Hopkins hospitals. The Center has created and maintains a large longitudinal database and biorepository (serum, plasma, DNA, RNA, peripheral blood mononuclear cells; tissue bank: skin, cancer, heart, lung, kidney, GI and other body tissue) of patients with scleroderma and other autoimmune diseases, along with appropriate controls. This is a dynamic entry registry with ongoing enrollment.

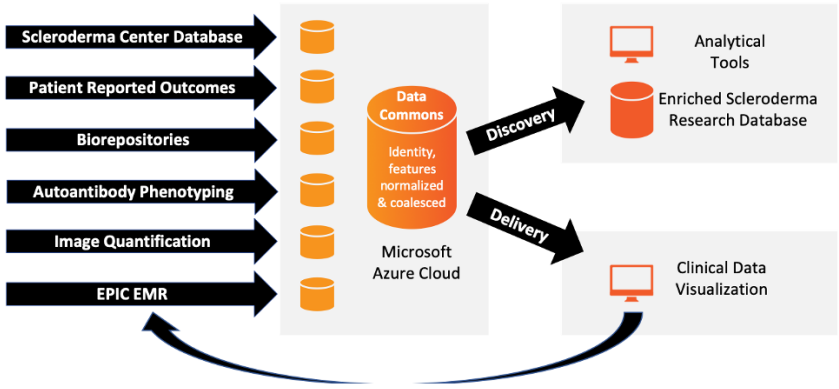


Figure 4.1: Data flow in the Johns Hopkins Precision Medicine Analytics Platform (PMAP)

Data from center participants have been ingested into the Johns Hopkins Precision Medicine Analytics Platform (PMAP). PMAP is an information technology component of the "Hopkins inHealth" Precision Medicine Program designed to enable learning within the practice of medicine. The Epic electronic health record at Johns Hopkins Medicine captures all transactions that comprise clinical care: clinical visits (history and examination, vital signs), laboratory measurements, prescriptions, procedures, and much more. Each evening, the main Epic data tables from the past 24 hours are copied to PMAP, an Azure cloud-based system for integrating and analyzing patient data from multiple sources. The major goal of Hopkins inHealth is to use modern measurement, data science, and connectivity tools to discover clinically relevant subgroups at scale and to deliver what we learn to impact the precision and value of health care. This study, a component of Hopkins inHealth, uses the nightly-updated data from PMAP as input to our patient-level data visualizations and predictions.

4.2.2 Steps to improve patient care

4.2.2.1 Approach

In the current practice of medicine, a clinician has access to historical and current data only about the patient at hand. That information is not typically organized or presented in a fashion for the clinician to appreciate the current status relative to its past. In addition, the patient's data are not placed within the context of other similar patients. For example outcomes for prior similar patients are not typically available. Clinicians therefore are forced to make qualitative judgements about the patient's status, trajectory, and likely benefits of different treatments, not fully informed by either the patient's own data or the experiences for other similar patients.

To improve patient care, this project takes the first steps towards:

- synthesizing and visualizing all of the historical and current data about a patient for clinician and patient use
- picturing the patient's current status and trajectory within the context of similar data for a user-specified sub-population of patients.

4.2.2.2 Delivering information to clinicians and patients

In order to improve patient care, a data science tool must ultimately be embedded within the clinical workflow used by physicians to guide their interaction with patients. We have already reached this end goal for a part of the work described here. Our visualization and analysis (VA) tool was initially developed in prototype form as an R Shiny application (App) (Chang et al., 2020). R Shiny

is a package that builds interactive web apps from R (R Core Team, 2020). Expert clinicians selected the key clinical information to be displayed. They reviewed and approved preliminary versions of all displays. As the next step, The Johns Hopkins Medicine Technology Innovation Center (TIC) used our R Shiny VA to implement within Epic, the JHM electronic health record, a version of our tool that physicians can directly use to test its value in clinical care. The TIC used the first clinician-approved version of part of our VA to build an Epic version called "Patient InSight". That version will be replaced by updated versions that incorporate estimation of patient trajectories, predictions of future trajectories and major events, and likely benefits of treatment options, once approved for clinical use. At the moment, the dynamic estimation and prediction of individuals' disease state and trajectory is only developed in the R Shiny app but will soon be embedded in the web-based VA. The final step necessary to improve patient care is to scale the use of the tool across all scleroderma and similar autoimmune disease patients at Johns Hopkins and beyond. That will require investment of a production scale version of the app by the Chief Medical Information Officer who prioritizes this app against others. Dissemination beyond JHM will be the domain of Johns Hopkins Technology Ventures. Our work is regularly shared with both of those offices so they are kept appraised of tool improvements in development.

4.2.3 Clinical data visualization

The VA illustrates a patient's aggregate clinical phenotype in a snapshot view, including cumulative disease manifestations, disease onset dates and autoantibody status. Any history of the following features are listed as disease manifestations: interstitial lung disease, pulmonary arterial hypertension, renal crisis, tendon friction rubs (TFRs), synovitis, myopathy, calcinosis, and other components of the 2013 American College of Rheumatology classification criteria for SSc (van den Hoogen et al., 2013). Comorbid conditions such as peripheral artery disease (PAD), coronary artery disease (CAD), atherosclerotic cardiovascular disease (ASCVD), hypertension (HTN), and cancer are also captured.

Longitudinal data are illustrated across multiple organ systems including: 1) cardiac (left ventricular ejection fraction (EF), right ventricular systolic pressure (RVSP), and right heart catheterization data), 2) pulmonary (percent predicted forced vital capacity – pFVC and diffusing capacity – pDLCO), 3) cutaneous (modified Rodnan skin score – mRSS), 4) gastrointestinal (Medsger GI severity scores relative to body mass index), 5) peripheral vasculature (Medsger Raynaud's scores capturing damage including digital pits, ulcerations and gangrene), and 6) muscle (proximal muscle strength on a 0-5 scale).

Additionally, we incorporated longitudinal body mass index (BMI), patient reported Health Assessment Questionnaire (HAQ) Disability Index (DI), organ involvement characterized by critical events, medications, additional laboratory data, and a summary of cardiopulmonary comorbid conditions. Longitudinal immunosuppressive medication exposure data is shown to assess

whether drug exposure alters trajectory in cutaneous and pulmonary parameters. Critical events are defined as longitudinal observations in multiple organs exceeding or falling below pre-specified thresholds. We indicate events for heart (EF < 50), PH (RVSP ≥ 45 and mean pulmonary arterial pressure (PAP) $\geq$ 25 for patients with right heart catheterization (RHC) data), lung (pFVC < 70 or pFVC < 60 and maximum mean PAP < 25 for patients with RHC), muscle (severity score of 4), GI (severity score of 4), and renal crisis by plotting them on a single time scale starting from scleroderma onset.

4.2.4 Estimation of disease state

4.2.4.1 Comparing individuals' trajectory to a user-defined subgroup

The tool incorporates the 10th, 50th and 90th percentile values for the entire Hopkins scleroderma cohort as a reference group for the pulmonary and cutaneous trajectories in particular. By plotting individuals' trajectories on top of the three reference lines, we get a better sense of patients' disease progression compared to others in their cohort. Moreover, we compare patients' trajectory to a user-specified subgroup based on clinical and demographic characteristics. This makes it easier for clinicians and patients to monitor their disease progression relative to a group of similar patients based upon known risk factors. In this application, we provide the group-specific quantile lines by filtering patients based on their biological sex, range for reference age of scleroderma onset, race, cutaneous subtype, and autoantibody status (positive for ACA, Scl-70, and RNAPol, or any combinations thereof).

4.2.4.2 Estimation of individuals' disease state and trajectory

A patient's true disease state or trajectory is an unobserved construct reflected in their longitudinal measurements and occurrences of sentinel events. We recognize that the disease status in multiple organs is measured longitudinally with error and often times missing. We use a tool that effectively estimates patients' disease state and rate of progression at any given moment and then communicate the estimates by integrating them within the visualization app. By fully utilizing information in multiple longitudinal markers, we maximize the efficiency of estimates of patient-specific and population trajectories.

We model the latent health state for lung function measured pFVC and pDLCO, heart function measured by RVSP and EF, and skin involvement measured by mRSS since individuals' disease onset. As is the clinical tradition, disease onset is defined by the earlier of the onset of Raynaud's phenomenon and first non-Raynaud's symptom. Prior to analysis, all 5 outcome measures are preprocessed using quantile normalization. Let Y_k be a vector of the observed values from each measure $k = 1, ..., 5$. Conceptually, the quantile normalized vector for each k is obtained by $\hat{\Phi}^{-1} \circ \hat{G}_k(Y_k)$, where $\hat{G}_k$ is an estimated distribution of the vector Y_k and $\hat{\Phi}^{-1}$ is the inverse of the standard normal distribution. Lastly, RVSP and mRSS observations are transformed by multiplying -1 so that increase/decrease in all 5 measures indicates better/worse health status.

We fit a Bayesian multivariate linear mixed effects model (MLMM) that accommodates the nested structure of the data: measures within time within a patient within a population. For each outcome measure, we select a set

of regression predictors (covariates) for the fixed effects. Here, we use age of onset, race, biological sex, cutaneous subtype, and indicators of positive ACA, RNAPol and Scl-70 antibodies. To model changes in patients' disease trajectories in time since onset, we also include a smooth function of time using natural splines with a degrees of freedom chosen based upon prior knowledge about over what time scale the disease state changes. Here we use 3 degrees of freedom in order to focus on the longer-term changes that are of greatest clinical relevance. The internal knots are placed at 10 and 30 years since onset; boundary knots at 0 and 40 years since onset. For patient specific random effects, we fit a random slope and intercept and two linear splines at 3 and 10 years from the last observation.

We fit the model using an R package **MCMCglmm** (Hadfield, 2010). For the fixed effects of both models, we use a diffuse independent normal prior centered around zero with a large variance (10^8). Weakly informative inverse-Wishart priors are placed on random effects and residual covariance matrices. Based upon prior knowledge of heterogeneity among patients, we set the prior distribution of the random intercepts to have mode one and random slopes to have mode of 0.005, with 20 degrees of freedom. The prior distribution of the residual covariance matrix is assumed to have the mode one for each measure, with 5 degrees of freedom. The degrees of freedom are chosen to make the distributions as diffuse as possible while guaranteeing them to be proper inverse-Wishart distributions.

Using the estimates from the model, we obtain and plot disease trajectories for each patient for the clinically selected measures pFVC, RVSP, and EF. These

were chosen because they are most important for treatment decisions and because the registry has a sizeable sample of people with long histories for these variables. Note that the estimated trajectories are transformed back to the original scales in the VA in R Shiny App for better communication of the results.

4.2.4.3 Prediction of future risk of critical events

The estimated level and the trend of a patient's disease trajectories serve as an indicator of the risk of having extreme values of biomarkers in the near future. Although estimated smooth trajectories in multiple organs describes how individuals' scleroderma evolved over time, the presence or absence of extreme values in patients' longitudinal observations is also clinically important. Observations falling below or rising above a clinically set threshold often serve as surrogates for critical events, and such events require immediate medical attention sometimes followed by more invasive and higher risk interventions. For example, $EF \leq 35$ implies severe heart failure and patients are often treated with implantable cardioverter defibrillator (ICD) placement.

We extend the work in Section 4.2.4.1 by projecting individuals' health trajectory into the future, to predict their risk of having critical events defined as following: $EF < 50$ and $EF < 35$ (cardiomyopathy), $RVSP \geq 45$ and $RVSP \geq 50$ (PH), and $pFVC \leq 70$ and $pFVC \leq 60$ (ILD). For each patient, we calculate the probability of having each of the events in the next 6, 12, and 18 months from the most recent visit using the Cross-validated Sequential Prediction (CVSP) using Multivariate Longitudinal Data (MLD) method. For details of

the CVSP method, see Chapter 3.

4.2.5 Evaluation of value added

We propose to study the value of this tool in the clinic by conducting a qualitative research study using human factors analytic strategies and a randomized clinical trial. In this paper, we introduce a way of efficiently using available resources to assess whether a patient level, longitudinal data visualization deployed in the scleroderma center clinic improves efficiency of care, patients understanding of their disease, and shared medical decision making by:

- improving the efficiency of care in terms of data capture,
- improving the efficiency of care in terms of delivery of information to patients,
- improving physician satisfaction,
- improving patient satisfaction and comfort with decision making, and
- reducing patient decisional regret.

Our plan is to conduct this study in two phases: 1) to assess the usability of the tool from the designer, provider and patient perspective, 2) to assess shared decision making from the patient perspective. Scleroderma patients who have previously consented to participate in our scleroderma center registry and who have at least one year of data will be included, as will scleroderma center providers. Detailed study procedures are described for each of these study populations in section 4.3.5.

4.3 Results

4.3.1 The Visualization Application

In Figure 4.2, we illustrate a patient's aggregate clinical phenotype and longitudinal data in a snapshot view. The selected tab shows the patient's cardiac data, which are EF and RVSP measurements over time. Regions indicating the severity of disease state are indicated by red (severe) and pink (mild) for each measure. Data from RHC are displayed for patients who had the procedure at the bottom of the tab. Below the longitudinal display of two heart measures, we show longitudinal display of any immunosuppressive medication used by the patient over time. The same plot of medications is also displayed in the lung and skin tab so that the medication history can be in the same view with the patient's longitudinal data of lung or skin.

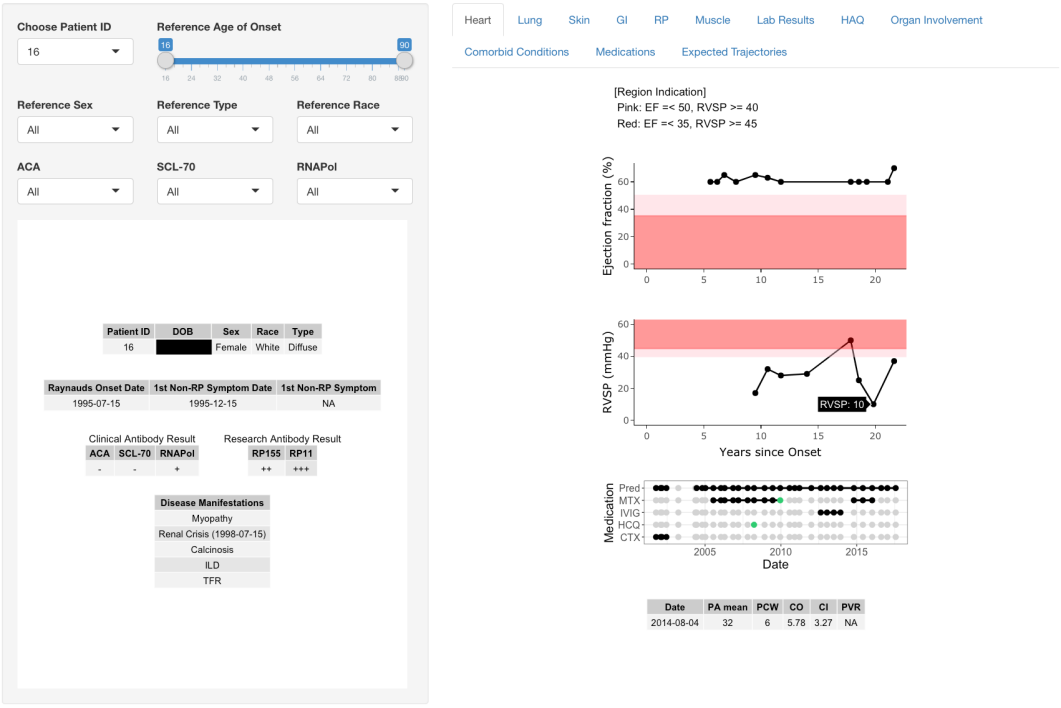


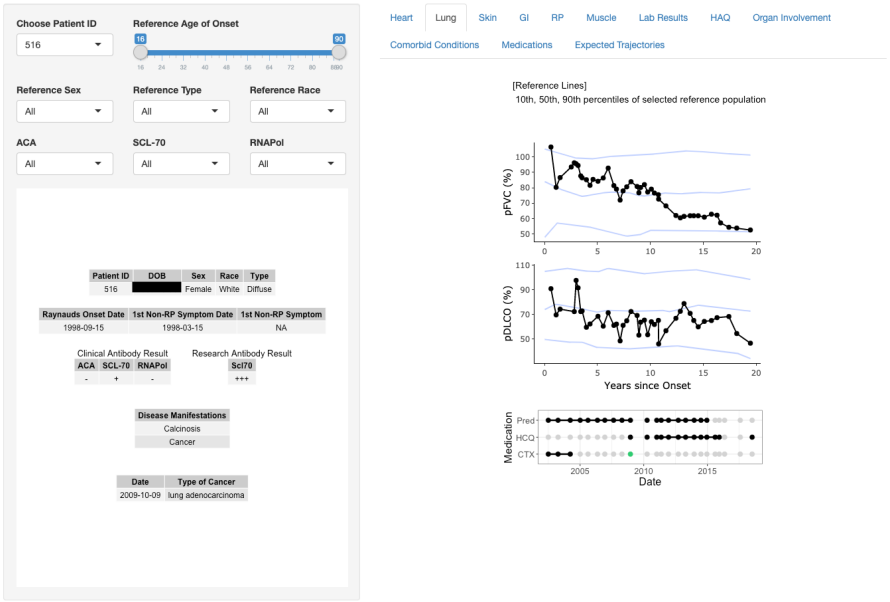
Figure 4.2: A patient's longitudinal observations. Users can view values for each points by hovering over the points in the graphs.

In the medication plot, the colors and connections between the points represent the patterns of medication exposure. Black points indicate that the patient is currently on medication, green indicates the patient is currently not on medication, but was on medication less than 6 months prior to visit. Grey points indicate the patient is not on medication and the lines are connecting black and green points if less than 15 months apart. Such method is employed to extrapolate the true medication exposure of the patients. The method is informative in that it separates missing medication data (no point plotted) from no exposure (grey), prior exposure only (green), and current exposure (black).

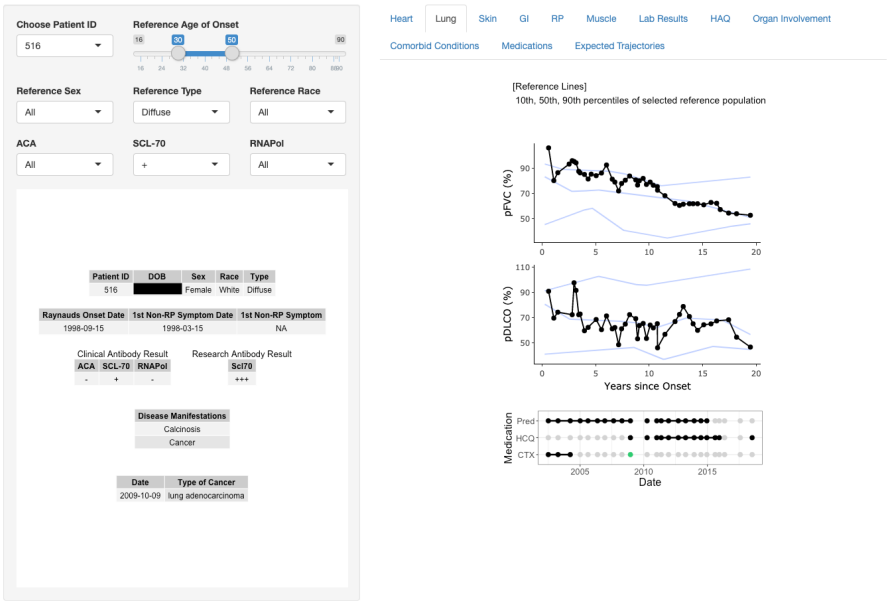
4.3.2 Trajectory within a reference population

In this section, we demonstrate the utility of presenting an individual patient's data relative to specific subgroup characteristics. In figure 4.3, the trajectory of a white woman who developed diffuse scleroderma with anti-topoisomerase 1 (Scl-70) antibodies with onset age of 40 years is compared to other patients with similar demographic and clinical characteristics, providing insight into how individual and combinations of risk factors may modify a patient's likely trajectory and outcome.

On the top panel (Figure 4.3 panel (a)), this patient's trajectory is compared with that of the overall scleroderma center cohort. On the bottom panel (panel (b)), that same patient's trajectory is compared with a reference population that is similar to the patient (onset age of 30 to 50 years, diffuse type, Scl-70 antibody positive). These data illustrate how the reference population changes the interpretation and one's perspective. When comparing this patient to the overall scleroderma population, we observe that her pFVC trajectory declines from the 90th percentile to 10th over 20 years of follow up, dropping rapidly below the 50th percentile line after 10 years. Relative to other similar patients, however, we see that her pFVC trajectory is better than that typically expected for the first 10 years of follow up and around the median afterwards.



(a) Lung trajectories of a patient and 10th, median, and 90th percentiles reference lines of the overall scleroderma population



(b) Lung trajectories of the same patient and 10th, median, and 90th percentiles reference lines of selected subpopulation

Figure 4.3: Screenshots of the Lung tab of the R Shiny app

4.3.3 The Web-based Visualization Application

In Figure 4.4, we present screenshots of the Epic-based visualization app called Patient InSight that reproduces the interface of the VA created in R Shiny App as demonstrated in previous sections. The information that displayed in each tab in the R Shiny App are presented in a single page view that users can scroll through. Users can select variables of interest to only view longitudinal data of the selected variables. Filtering of the reference population and the 10th, 50th, and 90th percentile reference lines are also featured. The application is currently used by physicians in the clinic mainly for the purpose of effectively scanning through patient data to assess patients' past and current health status prior to patients' visits. By implementing this version within Epic, we can now conduct the clinical studies discussed below. This was not possible using the Shiny app alone.

Figure 4.4: Screenshots of visualization assistance interface in Epic

4.3.4 Estimating a patient's risk of critical events

One feature that has not yet been implemented in the Epic-based VA is the estimation of the latent disease trajectory across multiple organ systems and the associated risk of critical events defined by specific thresholds. For each patient, we estimate the underlying disease trajectory for pFVC, EF, and RVSP through the most recent observation, and display them with the 95% prediction interval. An example is shown in Figure 4.5. Note that the predicted curves are obtained by jointly modeling the multiple measures. The method is particularly useful when some measures have fewer data points observed compared to others. In our case, patients generally have fewer observations for cardiac measures (EF, RVSP) and richer data for the pulmonary measures (pFVC, pDLCO). When there are only sparse data observed for a measure or a patient, we borrow strength from the other measures and from the entire cohort to produce more accurate and precise estimates.

In Figure 4.5, the estimated risk of 6 critical events (3 outcomes each with 2 severity levels) in the next 6 month are displayed. Projection time can be switched to 12 or 18 months from the drop-down menu at the top. As the events are directly defined by the value of the longitudinal measures, we project the estimated trajectory forward in time and calculate the risk of the future events using the estimated uncertainty around the prediction. The areas of the shaded regions in the graphs (lighter and darker shades of green and pink) indicate the estimated risk of having each event. The estimated risks are also quantified and tabulated next to the graphs. This particular patient has a high risk of having a clinically significant restrictive ventilatory defect

(93% for pFVC ≤ 70) and only a slight chance of having cardiomyopathy or PH (1% for EF < 50 and 3% for RVSP ≥ 45). The color coding is specific to the measure throughout the app where blue represents the lung and red the heart.

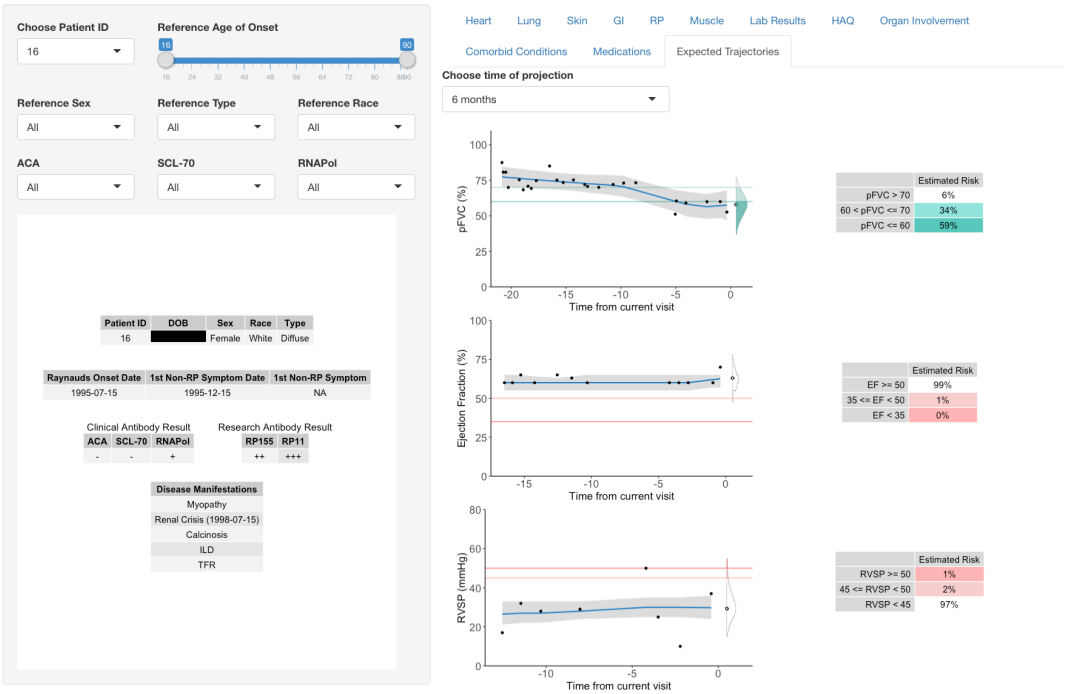


Figure 4.5: Estimated trajectories and risks for an individual patient

4.3.5 Evaluation Study

4.3.5.1 Assessing the utility of the VA

In this section, we present a two-phase study design to evaluate the utility of our visualization app.

Phase 1 - In this phase, the goal is to identify whether the VA improves the communication of information to patients, patient satisfaction, and provider satisfaction. We will first conduct semi-structured interviews for the 4 designers of the visualization tool (2 physicians, 2 biostatisticians) to capture

factors that influenced the design process. The tool will then be tested among providers seeing patients in the scleroderma center. Baseline information will be captured from study participants including age, sex, percentage of clinical time, number of years since rheumatology training or year of training, frequency of Epic EMR use, hours/week spent on a computer, computer platform most often used (PC or Mac), and general comfort with technology. Providers will also be asked how they typically prepare for seeing patients, how much time goes into data collection, which variables providers tend to share with patients, how they present evidence/data to patients, what information they would find valuable to have access to quickly, and to what extent they like making decisions independently vs with the input of their patients.

After these interviews are conducted, providers will have a session introducing them to the visualization tool, and how filters may be used to alter the scleroderma reference populations illustrated. Testing will be performed in the context of 3-4 real patient case scenarios; testing will be recorded (audio and video) and may be conducted over Zoom screen sharing with a study team member if clinic closures are prolonged with the COVID-19 pandemic. The cases will be selected to reflect typical issues arising the scleroderma center clinic – a) whether immunosuppressive therapy needs to be initiated or changed to treat one or more organ systems of involvement and b) whether more extensive and potentially invasive testing is needed to assess for scleroderma complications. Providers will be asked to assess each case using two methods – standard of care (e.g. examining the Epic EMR) and the visualization tool. Half of the providers will be asked to do the EMR assessments

first, and half will be asked to use the visualization tool first. Providers will be instructed to verbally describe their thought process during the session.

Throughout this process, we will record the time spent looking for distinct data elements, the types and numbers of tests or clinical features examined or not, and the number of clicks used. Total time examining the case before arriving at a decision will be noted. Providers will be asked about which decision they made and the level of confidence they have that they made a correct decision. Providers will also fill out a System Usability Scale questionnaire (Brooke, 1995, see Appendix B.1) to assess the value of the data visualization tool. Qualitative feedback will be solicited from providers at the end of the session to understand potential areas for improvement, alternative design ideas, whether any data elements were unclear, and user perception of task ease/difficulty. A study team member will take observation notes during the exercise to capture these factors and the sequence of tasks performed by providers.

After deployment of the visualization for at least 3 months, we will assess whether the providers are better able to see concerning trends or new patterns of disease evolution (i.e. higher rate of skin change coincides with worsening cardiac disease) than prior to use of the tool. Additionally, we will assess if this tool has altered the provider's work flow or practice habits in any way. Furthermore, we will assess whether the presentation of the data in this manner contributed to the generation of new scientific questions, and if so, what observations led to the generation of the new ideas. These assessments will occur twice over the duration of the yearlong study.

For patient participants, we will perform a qualitative evaluation using semi-structured interviews to gain early feedback on a patient's perception of seeing his/her longitudinal trajectory illustrated. This may be conducted over Zoom in the setting of COVID-19. In particular we will solicit feedback on how patients feel about seeing their data illustrated relative to aggregate cohort data, and how they understand and interpret the data in the tool. We will recruit up to 15 patients in this phase of the study. Feedback gained may inform whether modifications need to be made to the visualization tool and whether any additional factors need to be studied in Phase 2 detailed below. See Appendix B.2 for the questionnaires that will be used in Phase 1.

Phase 2 - The study population for this phase will be scleroderma center patient participants. Through our approved registry IRB, cohort patients will already have completed the Combined Outcome Measure for Risk Communication and Treatment Decision Making Effectiveness (COMRADE) questionnaire (Edwards et al., 2003, see Appendix B.3); this patient-reported outcome measure assesses how patients feel about transmission of information to make medical decisions, satisfaction with decision making, and decisional regret.

Through this new proposal, patients will be consented to participate in study of the visualization tool. Patients will be randomized 1:1 to deployment of the visualization tool – that is, half will be exposed to the visualization and half will not. This randomization will enable us to determine whether the visualization tool improves patients' understanding of their disease and sense of shared decision making. After the clinical visit, patients will be asked to rate their understanding of their disease state and complete the COMRADE

questionnaire. Patients who are randomized to the visualization tool will complete an additional questionnaire (see Appendix B.4) specific to the tool's usability. Separate study visits are not required for this phase of the study. All data will be collected in the context of routine clinical care or telemedicine visits.

4.3.5.2 Statistical analysis

In this section, we lay out the statistical analysis plan using data captured from the Phase 2 questionnaire (see Appendix B.4) and estimate adequate sample size. Our primary goal is to compare the satisfaction scores for two patient groups: with visualization assistance (VA) and without VA. Within each provider, an equal number of patients will be randomized into the two groups.

We plan to fit a series of linear random effects models, where the outcome variable is a satisfaction score standardized to have mean 0 and variance 1. For the most basic model, we include a patient group indicator as the fixed effects, and provider indicator as the random effects. By fitting the random effects model, we control for the correlation of observations within a provider. As the secondary analysis, we will additionally control for baseline covariates including patient's age, years since onset, years since their first visit at the clinic, and disease severity score. We will also estimate treatment effects separately for each subgroup of patients defined by the covariates. By comparing the estimated treatments to that of the base model and also by comparing the treatment effects across the subgroups, we can check if the

selected covariates confound or modify the treatment effects.

After estimating the standard deviation (σ) of the outcome variables from the model, we can determine more precisely the number of patients required to achieve a given power to detect a group mean difference or the effect size. To achieve a power (β) of at least 0.8 and $\alpha = 0.05$, we require around 64 patients in each group to detect an effect size of 0.5 standard deviations assuming σ=1. Effect size of 0.3 requires 175 patients per group, and effect size of 0.7 requires 33 patients per group with the same conditions. With our design of 100 patients total, we have 80% power to detect the difference in the satisfaction scores of the standard of care and visualization tool.

4.4 Discussion

This visualization and analysis tool is designed to improve both provider and patient satisfaction, improve clinical decision making, and reduce decisional regret for patients with this complex rheumatic disease. We also present a way of testing the utility of our visualization tool to improve shared medical decision making and patients' understanding of their disease state. In addition, we seek to test whether the tool improves the efficiency of clinical care.

Development of the visualization tool and testing its utility required collective effort of clinicians, statisticians, Johns Hopkins TIC staff, and human factors specialists. While we plan to carry out all components of Phase 1 evaluation by the end of 2020, continuous efforts are made to improve the VA. In particular, we are in the process of testing and calibrating the tools that estimate and project health trajectories in order for them to be embedded in

the Epic Patient Tool and eventually be useful in clinic.

One of the major goals for statisticians is to estimate the medication effect based on patients' health trajectory and medication history. We are currently building a Bayesian causal effects model that yields predictions about the likely effects of selected interventions, for a specific patient, or in a population of patients. Embedding patients' future projections and the changes thereof in trajectory for different combinations of medications is the final goal of this project.

Another long-term goal of this project is disseminating this tool as a resource beyond Johns Hopkins Scleroderma Center. The visualization app and statistical models used in this study allow flexible parameterization and can be applied to display clinical data and model health trajectories for other complex diseases. In particular, we saw an opportunity to leverage the resources of the Hopkins inHealth initiative to build a framework that could be scaled up and generalizable for multiple studies and clinical applications.

Chapter 5

Discussion and Conclusion

The works presented in this thesis comprise methods to better utilize available clinical data to improve clinical care for scleroderma. In our application, a patient's health state is reflected in multiple irregularly spaced longitudinal measures and events. In some instances, the events of interest represent threshold crossings of those same longitudinal measures. A first objective was to estimate smooth individual and population health trajectories across different organ systems using noisy and, for many individuals, sparse data. We estimated the trajectories by selecting and estimating multivariate Bayesian hierarchical models that accommodate the nested structure of the observed outcome variables and multiple clinically-relevant predictor variables. From the model, we furthered our understanding of the complexity of the disease by studying estimated disease progression in multidimension space for clinically defined subpopulations as well as the estimated correlations across measures and time.

To achieve the clinical aim, we confronted the question: in cases like scleroderma, does fitting a more complex multivariate hierarchical model

("combined model") produce substantially more efficient estimates compared to fitting a set of "separated models", that is, separate univariate models for each measure. In regression analysis, this question was raised by Zellner (Zellner, 1962) who coined the term "seemingly unrelated regression" equations or SUR. He showed that the coefficient estimation using the GLS (Aitken, 1934) is asymptotically more efficient compared to the OLS, and that the efficiency increases as the error terms from different equations become more cross-correlated and as the predictor variables in different equations become less correlated. The OLS estimates are only as efficient as the GLS, when each equations system representing the relationship between each outcome variable and its set of predictor variables are uncorrelated in random error terms across outcomes. But multivariate linear mixed models, which additionally involves random effects defined through time across all outcomes, are not separable into individual equation systems without efficiency loss as can happen in SUR where the equations are connected only through random error terms. With mixed effect models, there is also an additional question: how does the seemingly unrelated approach (separated models) do for Bayes estimates of individual slopes? All previous work was about the fixed effects. We derived a set of generalizable formulae to compare the relative efficiency of population and individual-level estimates from the fully efficient combined model and the simpler separated models. The relative performance of the combined model depended on the amount of the available data and the degree of correlation between the measure of interest and the other measures. For estimating the fixed effects in the scleroderma application, both models were provided with rich data, hence we observe minimal gain in reduced variance

by fitting the combined model. On the other hand, there are sizeable gains in estimating random effects for those individual's for whom the relative number of observations in the measure of interest is smaller than those in other correlated measures. The degree of efficiency gain increases with the degree of correlation.

We found that the reduction in MSE mostly results from reduced bias. For individuals who have only a few data points available for a given measure, the data for the measure alone cannot accurately reflect the underlying disease state of the individual. Hence, fitting the separated models results in greater shrinkage towards the measure-specific mean and results in larger bias. The bias is reduced when fitting the combined model, where the random effects estimator borrows strength from data-rich measures and yields more efficient estimates.

Estimating the risk of future clinical events is another clinical priority, so we built prediction algorithms that fully utilize the information in patients' past trajectories. Logistic regression models and machine learning algorithms such as ensemble methods using decision trees are frequently used to build prediction models where the outcome of interest is a binary event. However, it is difficult to incorporate the multidimensional trajectory information in a disease like scleroderma for such models. Also, there is a huge loss of information by transforming a continuous outcome into a binary variable. Hence, we used the framework developed to estimate individual trajectory by jointly modeling multiple markers and extended the models to predicting patients' risk of having critical events in the near future.

We implemented the prediction calculations by developing a cross-validated, sequential prediction algorithm (CVSP) for multivariate longitudinal data (MLD). The algorithm sequentially produces the most likely trajectory and the risk of clinical events as additional data points are observed for a patient. The predictions are made without refitting the model to incorporate new observations for a patient using K-fold cross-validation method. We divide patients into k folds of data. To obtain predictions for a patient in the kth fold, we produce estimates of the random error and random effects covariance matrices leaving out the kth fold. In this way the estimated matrices by construction are independent of the left out data. We then use the fixed and random effect estimates which are functionals of the estimated covariance matrices to sequentially estimate a patient's risk of events given only the past data. The CVSP for MLD produces predictions of six clinical outcomes of interest with substantially higher precision as compared to the empirical prediction methods using logistic regression models. Moreover, we showed that CVSP increases in precision as more data are observed for a given patient and that, even with no observations for an individual's measure, CVSP yields predictions with considerable precision. The result implies that we can still rely on the estimation of the risk of an event, when there is no available data for the corresponding measure, because our model borrows strength from the patient's data in other measures and also from other patients with similar characteristics. It should be noted that the events used in our application are directly determined by longitudinal measures used in the model. Modeling for events that are not functions of the longitudinal measures, i.e. death, requires additional modeling work.

In Chapter 4, we presented an interactive visualization application for physicians and/or patients that displays a patient's longitudinal data, estimates of their prior trajectory with its uncertainty, and prediction of future measures and the associated risks of critical events. The tool is designed to automatically aggregate clinical phenotype and longitudinal data in a snapshot view to facilitate evidence-based practice and shared decision making for physicians and help patients understand their disease status better. There are multiple challenges in order for the visualization tool to be used in the clinic. First, we need to test the usability of the tool, demonstrating improved satisfaction of both physicians and patients. We designed the qualitative studies and trials in two phases and throughout will continue to improve the tool based on the feedback we gather from the questionnaires. Currently, the Epic-based tool does not include the estimation and prediction tools mentioned above. One of the major future goals of this project is to further update the tool with the predictions and design trials to test its usability for its broader use.

The methods developed in this thesis have limitations despite our effort to give them broad domains of application. Each limitation represents an opportunity for further work. First, we rely upon the multivariate Gaussian distribution in order to derive estimates and predictions about an individual's trajectory. We accommodate non-Gaussian marginals by quantile normalization and further check for violations of multivariate Gaussian random effects and residuals. Nevertheless, the Gaussian assumption makes the methods sensitive to outliers in either the random effects or errors. Model checking

is therefore essential with current version. We plan to broaden the application by allowing the random effects and errors to follow a t-distribution with smaller degrees of freedom rather than the Gaussian to admit longer-tailed distributions.

Second, our clinician colleagues believe strongly that there are subsets of patients over and above those represented by fixed effects such as auto-antibody type. To accommodate this possibility, we can extend the model by allowing the random effects distributions to be a mixture of multiple Gaussians with different means. This approach has been used in "growth curve models" in the social sciences literature (Muthen, 2001).

Finally, we have made progress toward answering two of our motivating clinical questions asked by patients: (1) what is my current disease state; (2) what is my trajectory - where can I expect to be in the near future? The third question is: of the treatments currently available, which one is best for me going forward? The models proposed here have extensions to incorporate the causal question about treatments. The fixed effects regressions can include treatment indicators to quantify observed difference between treatment subgroups. The prior distributions for the treatment effects can incorporate the results of relevant clinical trials. Heterogeneity among patients in the effects of treatments can be included using random effects. However, treatments are rarely assigned through randomization in clinical practice. Hence, it is necessary to model the treatment assignment as another outcome in the multivariate model. With colleagues, we currently are working toward having useful answers for the third question and to communicate them to patients

and clinicians in the apps developed here and test them in the subsequent clinical studies as described in Chapter 4.

Appendix A

Supplementary materials for chapter 2

A.1 Mean Squared Error and bias-variance decomposition

The mean squared error (MSE) of an estimator $\theta \in \mathbf{R}^d$ is defined as

$$E(||\hat{\theta} - \theta||^2) = E(\sum_{j=1}^{d}(\hat{\theta}_j - \theta_j)^2) = Tr(Var(\hat{\theta})) + ||Bias(\hat{\theta})||^2$$

where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. Note that $Var(\hat{\theta})$ is the covariance matrix of $\hat{\theta}$ and its trace is $\sum_{j=1}^{d} Var(\hat{\theta}_j)$. Since $MSE(\hat{\theta}, \theta) = \sum_{j=1}^{d} E((\hat{\theta}_j - \theta_j)^2)$, it is sufficient to show $E((\hat{\theta} - \theta)^2) = Var(\hat{\theta}) + Bias^2(\hat{\theta})$ to prove the above result.

$$E((\hat{\theta}-\theta)^2) = E((\hat{\theta}-E(\hat{\theta}))+(E(\hat{\theta})-\theta))^2 = E\{(\hat{\theta}-E(\hat{\theta}))^2+(E(\hat{\theta})-\theta)^2$$

$$+2(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)\}$$

$$= E(\hat{\theta}-E(\hat{\theta}))^2+(E(\hat{\theta})-\theta)^2+2(E(\hat{\theta})-E(\hat{\theta}))(E(\hat{\theta})-\theta)\}$$

$$= E(\hat{\theta}-E(\hat{\theta}))^2+(E(\hat{\theta})-\theta)^2$$

$$= Var(\hat{\theta})+Bias^2(\hat{\theta})$$

Under our assumptions, $E(Y) = X\beta$ and $var(Y) = V_C$, the fixed effect estimates $\hat{\beta}_S$ and $\hat{\beta}_C$ are unbiased as

$$E(\hat{\beta}_S) = (X^T W_S X)^{-1} X^T W_S E(Y) = \beta$$

$$E(\hat{\beta}_C) = (X^T W_C X)^{-1} X^T W_C E(Y) = \beta$$

and

$$MSE(\hat{\beta}_S, \beta) = Tr(var(\hat{\beta}_S)) = Tr((X^T W_S X)^{-1} X^T W_S var(Y) W_S X (X^T W_S X)^{-1})$$

$$= Tr((X^T W_S X)^{-1} X^T W_S V_C W_S X (X^T W_S X)^{-1})$$

$$MSE(\hat{\beta}_C, \beta) = Tr(var(\hat{\beta}_C)) = Tr((X^T W_C X)^{-1})$$

MSE for random effects $\hat{b}_{Si}$ and $\hat{b}_{Ci}$ can be decomposed into variance and bias components. Note that under our assumptions, $var(\epsilon_i) = \Sigma_{Ci}$ and $var(b_i) = D_C$. For the separated model,

$$E_{b_i}\{MSE(\hat{b}_{Si}, b_i)\} = E_{b_i}\{Tr(var_{\hat{b}_{Si}|b_i}(\hat{b}_{Si}|b_i))\} + E_{b_i}\{||Bias(\hat{b}_{Si})||^2\}$$

where the variance component is

$$E_{b_i}\{Tr(var_{y_i|b_i}(\hat{b}_{Si}|b_i))\} = Tr\{D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si})\Sigma_{Ci}$$

$$\times (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i D_S + X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} V_{Cn} W_{Sn} X_n\}$$

since $var_{\hat{b}_{Si}|b_i}(\hat{b}_{Si}|b_i) = var_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(y_i - X_i\hat{\beta}_S)|b_i)$

$$= var_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(y_i - X_i(X^T W_S X)^{-1} \sum_{n=1}^{m} X_n^T W_{Sn} y_n)|b_i)$$

$$= var_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(y_i - X_i(X^T W_S X)^{-1} X_i^T W_{Si} y_i - X_i(X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} y_n)|b_i)$$

$$= var_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(I - X_i(X^T W_S X)^{-1} X_i^T W_{Si})\epsilon_i)$$

$$+ var_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si} X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} y_n)$$

$$= D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si})\Sigma_{Ci}(W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i D_S$$

$$+ D_S Z_i^T W_{Si} X_i (X^T W_S X)^{-1} (\sum_{n \neq i}^{m} X_n^T W_{Sn} V_{Cn} W_{Sn} X_n)(D_S Z_i^T W_{Si} X_i (X^T W_S X)^{-1})^T$$

The bias component is estimated as following

$$E_{b_i}\{||Bias(\hat{b}_{Si})||^2\} = E_{b_i}\{||(E(\hat{b}_{Si}|b_i) - b_i)||^2\} = E_{b_i}\{Tr((E(\hat{b}_{Si}|b_i) - b_i)(E(\hat{b}_{Si}|b_i) - b_i)^T)\}$$

$$= Tr\{(D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I) E_{b_i}(b_i b_i^T)$$

$$\times (D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I)^T\}$$

$$= Tr\{(D_S Z_i^T (W_{Si} - W_{Si}X_i(X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I) D_C$$

$$\times (D_S Z_i^T (W_{Si} - W_{Si} X_i (X^T W_S X)^{-1} X_i^T W_{Si}) Z_i - I)^T \}$$

since

$$E_{\hat{b}_{Si}|b_i}(\hat{b}_{Si}|b_i) = E_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(y_i - X_i \hat{\beta}_S)|b_i)$$

$$= E_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}(y_i - X_i (X^T W_S X)^{-1} \sum_{n=1}^{m} X_n^T W_{Sn} y_n)|b_i)$$

$$= E_{\hat{b}_{Si}|b_i}(D_S Z_i^T W_{Si}\{y_i - X_i (X^T W_S X)^{-1} X_i^T W_{Si} y_i$$

$$- X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} y_n\}|b_i)$$

$$= D_S Z_i^T W_{Si}\{(I - X_i (X^T W_S X)^{-1} X_i^T W_{Si})(X_i \beta + Z_i b_i)$$

$$- X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} X_n \beta\}$$

$$= D_S Z_i^T W_{Si}\{(I - X_i (X^T W_S X)^{-1} X_i^T W_{Si}) Z_i b_i$$

$$+ (I - X_i (X^T W_S X)^{-1} X_i^T W_{Si}) X_i \beta - X_i (X^T W_S X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Sn} X_n \beta\}$$

$$= D_S Z_i^T W_{Si}\{(I - X_i (X^T W_S X)^{-1} X_i^T W_{Si}) Z_i b_i$$

$$+ X_i \beta - X_i (X^T W_S X)^{-1} \sum_{n=1}^{m} X_n^T W_{Sn} X_n \beta\}$$

$$= D_S Z_i^T W_{Si}\{(I - X_i (X^T W_S X)^{-1} X_i^T W_{Si}) Z_i b_i$$

Similarly, for the combined model

$$E_{b_i}\{MSE(\hat{b}_{Ci}, b_i)\} = E_{b_i}\{Tr(var_{\hat{b}_{Ci}|b_i}(\hat{b}_{Ci}|b_i))\} + E_{b_i}\{||Bias(\hat{b}_{Ci})||^2\}$$

where the variance component is

$$E_{b_i}\{Tr(var_{y_i|b_i}(\hat{b}_{Ci}|b_i))\} = Tr\{D_C Z_i^T (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci})\Sigma_{Ci}$$

$$\times (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci}) Z_i D_C + X_i (X^T W_C X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Cn} X_n\}$$

since $var_{\hat{b}_{Ci}|b_i}(\hat{b}_{Ci}|b_i) = var_{\hat{b}_{Ci}|b_i}(D_C Z_i^T W_{Ci}(y_i - X_i\hat{\beta}_C)|b_i)$

$$= var_{\hat{b}_{Ci}|b_i}(D_C Z_i^T W_{Ci}(I - X_i(X^T W_C X)^{-1} X_i^T W_{Ci})\epsilon_i)$$

$$+ var_{\hat{b}_{Ci}|b_i}(D_C Z_i^T W_{Ci} X_i (X^T W_C X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Cn} y_n)$$

$$= D_C Z_i^T (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci})\Sigma_{Ci}(W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci}) Z_i D_C$$

$$+ D_C Z_i^T W_{Ci} X_i (X^T W_C X)^{-1} \sum_{n \neq i}^{m} X_n^T W_{Cn} X_n (D_C Z_i^T W_{Ci} X_i (X^T W_C X)^{-1})^T$$

The bias component is

$$E_{b_i}\{||Bias(\hat{b}_{Ci})||^2\} = E_{b_i}\{Tr((E(\hat{b}_{Ci}|b_i) - b_i)(E(\hat{b}_{Ci}|b_i) - b_i)^T)\}$$

$$= Tr\{(D_C Z_i^T (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci}) Z_i - I) D_C$$

$$\times (D_C Z_i^T (W_{Ci} - W_{Ci}X_i(X^T W_C X)^{-1} X_i^T W_{Ci}) Z_i - I)^T\}$$

since $E_{\hat{b}_{Ci}|b_i}(\hat{b}_{Ci}|b_i) = E_{\hat{b}_{Ci}|b_i}(D_C Z_i^T W_{Ci}(y_i - X_i\hat{\beta}_C)|b_i)$

$= D_C Z_i^T W_{Ci}\{(I - X_i(X^T W_C X)^{-1} X_i^T W_{Ci}) Z_i b_i.$

Lastly, MSE for $\hat{y}_{Si} = X_i\hat{\beta}_S + Z_i\hat{b}_i = X_i\hat{\beta}_S + Z_i D_S Z_i^T W_{Si}(y_i - X_i\hat{\beta}_S)$ can be written as

$$E_{b_i}\{MSE(\hat{y}_{Si},\, E(\hat{y}_i|b_i))\} = E_{b_i}\{Tr(var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i))\} + E_{b_i}\{||Bias(\hat{y}_{Si})||^2\}$$

$E_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i)$ and $var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i)$ is derived where

$M_{Si} = (X_i - Z_i D_S Z_i^T W_{Si} X_i)(X^T W_S X)^{-1}$

$$
\begin{aligned}
E_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i) &= E_{\hat{y}_{Si}|b_i}(X_i\hat{\beta}_S + Z_i D_S Z_i^T W_{Si}(y_i - X_i\hat{\beta}_S)) \\
&= E_{\hat{y}_{Si}|b_i}\{M_{Si}\sum_{n=1}^{m} X_n^T W_{Sn} y_n + Z_i D_S Z_i^T W_{Si} y_i\} \\
&= E_{\hat{y}_{Si}|b_i}\{M_{Si} X_i^T W_{Si} y_i + Z_i D_S Z_i^T W_{Si} y_i + M_i \sum_{n\neq i}^{m} X_n^T W_{Sn} y_n\} \\
&= M_{Si} X_i^T W_{Si}(X_i\beta + Z_i b_i) + Z_i D_S Z_i^T W_{Si}(X_i\beta + Z_i b_i) + M_{Si}\sum_{n\neq i}^{m} X_n^T W_{Sn} X_n \beta \\
&= X_i\beta + \{M_{Si} X_i^T + Z_i D_S Z_i^T\} W_{Si} Z_i b_i
\end{aligned}
$$

$$
\begin{aligned}
var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i) &= var_{\hat{y}_{Si}|b_i}\{M_{Si} X_i^T W_{Si} y_i + Z_i D_S Z_i^T W_{Si} y_i + M_{Si}\sum_{n\neq i}^{m} X_n^T W_{Sn} y_n\} \\
&= (M_{Si} X_i^T + Z_i D_S Z_i^T) W_{Si} \Sigma_{Ci} W_{Si}(M_{Si} X_i^T + Z_i D_S Z_i^T)^T \\
&+ M_{Si}(\sum_{n\neq i}^{m} X_n^T W_{Sn} V_{Cn} W_{Sn} X_n) M_{Si}^T
\end{aligned}
$$

Hence,

$$
\begin{aligned}
E_{b_i}\{Tr(var_{\hat{y}_{Si}|b_i}(\hat{y}_{Si}|b_i))\} = Tr((&M_{Si} X_i^T + Z_i D_S Z_i^T) W_{Si} \Sigma_{Ci} W_{Si}(M_{Si} X_i^T + Z_i D_S Z_i^T)^T \\
&+ M_{Si}(\sum_{n\neq i}^{m} X_n^T W_{Sn} V_{Cn} W_{Sn} X_n) M_{Si}^T)
\end{aligned}
$$

and

$$
E_{b_i}\{||Bias(\hat{y}_{Si})||^2\} = E_{b_i}\{Tr((E(\hat{y}_{Si}|b_i) - (X_i\beta + Z_i b_i))(E(\hat{y}_{Si}|b_i) - (X_i\beta + Z_i b_i))^T)\}
$$

$$= Tr(E_{b_i}\{(\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)Z_ib_ib_i^TZ_i(\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)^T\})$$

$$= Tr((\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)Z_iD_{Ci}Z_i(\{M_{Si}X_i^T + Z_iD_SZ_i^T\}W_{Si} - I)^T)$$

MSE for $\hat{y}_{Ci} = X_i\hat{\beta}_C + Z_iD_CZ_i^TW_{Ci}(y_i - X_i\hat{\beta}_C)$ is

$$E_{b_i}\{MSE(\hat{y}_{Ci},\, E(\hat{y}_i|b_i))\} = E_{b_i}\{Tr(var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i))\} + E_{b_i}\{||Bias(\hat{y}_{Ci})||^2\}$$

Letting $M_{Ci} = (X_i - Z_iD_CZ_i^TW_{Ci}X_i)(X^TW_CX)^{-1}$,

$$E_{\hat{y}_{Si}|b_i}(\hat{y}_{Ci}|b_i) = E_{\hat{y}_{Ci}|b_i}(X_i\hat{\beta}_C + Z_iD_CZ_i^TW_{Ci}(y_i - X_i\hat{\beta}_C))$$

$$= E_{\hat{y}_{Ci}|b_i}\{M_{Si}\sum_{n=1}^{m}X_n^TW_{Cn}y_n + Z_iD_CZ_i^TW_{Ci}y_i\}$$

$$= E_{\hat{y}_{Ci}|b_i}\{M_{Ci}X_i^TW_{Ci}y_i + Z_iD_CZ_i^TW_{Ci}y_i + M_i\sum_{n\neq i}^{m}X_n^TW_{Cn}y_n\}$$

$$= M_{Ci}X_i^TW_{Ci}(X_i\beta + Z_ib_i) + Z_iD_CZ_i^TW_{Ci}(X_i\beta + Z_ib_i) + M_{Ci}\sum_{n\neq i}^{m}X_n^TW_{Cn}X_n\beta$$

$$= X_i\beta + \{M_{Ci}X_i^T + Z_iD_CZ_i^T\}W_{Ci}Z_ib_i$$

$$var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i) = var_{\hat{y}_{Ci}|b_i}\{M_{Ci}X_i^TW_{Ci}y_i + Z_iD_CZ_i^TW_{Ci}y_i + M_{Ci}\sum_{n\neq i}^{m}X_n^TW_{Cn}y_n\}$$

$$= (M_{Ci}X_i^T + Z_iD_CZ_i^T)W_{Ci}\Sigma_{Ci}W_{Ci}(M_{Ci}X_i^T + Z_iD_CZ_i^T)^T$$

$$+ M_{Ci}(\sum_{n\neq i}^{m}X_n^TW_{Cn}X_n)M_{Ci}^T$$

Hence,

$$E_{b_i}\{Tr(var_{\hat{y}_{Ci}|b_i}(\hat{y}_{Ci}|b_i))\} = Tr((M_{Ci}X_i^T + Z_iD_CZ_i^T)W_{Ci}\Sigma_{Ci}W_{Ci}(M_{Ci}X_i^T + Z_iD_CZ_i^T)^T)$$

$$+M_{Ci}(\sum_{n\neq i}^{m} X_n^T W_{Cn} X_n) M_{Ci}^T$$

and

$$E_{b_i}\{||Bias(\hat{y}_{Ci})||^2\} = E_{b_i}\{Tr((E(\hat{y}_{Ci}|b_i) - (X_i\beta + Z_i b_i))(E(\hat{y}_{Ci}|b_i) - (X_i\beta + Z_i b_i))^T)\}$$

$$= Tr(E_{b_i}\{(\{M_{Ci}X_i^T + Z_i D_C Z_i^T\}W_{Ci} - I)Z_i b_i b_i^T Z_i(\{M_{Ci}X_i^T + Z_i D_C Z_i^T\}W_{Ci} - I)^T\})$$

$$= Tr((\{M_{Ci}X_i^T + Z_i D_C Z_i^T\}W_{Ci} - I)Z_i D_{Ci} Z_i(\{M_{Ci}X_i^T + Z_i D_C Z_i^T\}W_{Ci} - I)^T)$$

A.2 Correlation matrices with varying degrees of correlation across patient-specific trends

	pFVC	pDLCO	EF	RVSP	mRSS
pFVC	1.00	0.64	0.10	0.37	0.27
pDLCO	0.64	1.00	-0.10	0.40	0.20
EF	0.10	-0.10	1.00	0.10	-0.10
RVSP	0.37	0.40	0.10	1.00	0.13
mRSS	0.27	0.20	-0.10	0.13	1.00

Table A.1: Correlation across random slope components from $C_{0.1}^{'}$

	pFVC	pDLCO	EF	RVSP	mRSS
pFVC	1.00	0.64	0.20	0.37	0.27
pDLCO	0.64	1.00	-0.20	0.40	0.20
EF	0.20	-0.20	1.00	0.20	-0.20
RVSP	0.37	0.40	0.20	1.00	0.13
mRSS	0.27	0.20	-0.20	0.13	1.00

Table A.2: Correlation across random slope components from $C_{0.2}^{'}$

	pFVC	pDLCO	EF	RVSP	mRSS
pFVC	1.00	0.62	0.28	0.38	0.26
pDLCO	0.62	1.00	-0.28	0.39	0.21
EF	0.28	-0.28	1.00	0.29	-0.28
RVSP	0.38	0.39	0.29	1.00	0.12
mRSS	0.26	0.21	-0.28	0.12	1.00

Table A.3: Correlation across random slope components from $C'_{0.3}$

Appendix B

Supplementary materials for chapter 4

B.1 System Usability Scale questionnaire

1. I think that I would like to use the visualization frequently.

   Strongly disagree Disagree Neutral Agree Strongly Agree

2. I found the visualization unnecessarily complex.

   Strongly disagree Disagree Neutral Agree Strongly Agree

3. I thought the visualization was easy to use.

   Strongly disagree Disagree Neutral Agree Strongly Agree

4. I think that I would need the support of a technical person to be able to use this visualization.

   Strongly disagree Disagree Neutral Agree Strongly Agree

5. I found the various functions in this visualization were well integrated.

   Strongly disagree Disagree Neutral Agree Strongly Agree

6. I thought there was too much inconsistency in this visualization.

   Strongly disagree Disagree Neutral Agree Strongly Agree

7. I would imagine that most people would learn to use this visualization very quickly.

   Strongly disagree Disagree Neutral Agree Strongly Agree

8. I found the visualization very cumbersome to use.

   Strongly disagree Disagree Neutral Agree Strongly Agree

9. I felt very confident using the visualization.

   Strongly disagree Disagree Neutral Agree Strongly Agree

10. I needed to learn a lot of things before I could get going with this visualization.

    Strongly disagree Disagree Neutral Agree Strongly Agree

B.2 Provider and patient questionnaire for Phase 1

B.2.1 Patient questionnaire

Understanding your scleroderma

1) How well do you understand your scleroderma today?

   IF answer is positive, can you describe a little bit of your understanding?

   IF answer is negative, can you elaborate what you have understood? What you did not understand? What are the aspects that you may need for a better understanding? Why do you think you did not understand? What is your expectation of the understanding? Was the expectation met?

2) How well do you understand the trajectory of your scleroderma up to this point?

IF answer is positive, can you describe how your disease has changed/progressed over time?

IF answer is negative, can you elaborate on what changes you have noticed since you were first diagnosed? What changes did you expect? What changes did you not expect? Was the timing or rate of change expected?

3) How well do you understand the likely future trajectory or prognosis of your scleroderma?

IF answer is positive, do you expect your scleroderma to change/progress over the next year or five years? Can you describe how you expect your scleroderma to change/progress?

IF answer is negative, can you describe what changes in your scleroderma are expected in general? your expectation of disease progression over the next year/five years? What changes if any were explained as possibilities?

4) How well do you understand the reason you have been prescribed or not prescribed certain medications for your scleroderma?

   IF answer is positive, which medications are you taking for your scleroderma and why?

   IF answer is negative, what aspect of your scleroderma do you think requires medication? What areas of your scleroderma would you like medication for if there was one to treat a particular symptom?

   Is there additional information that would better help you understand your disease?

   IF answer is positive, what information would be helpful?

   IF answer is negative, what information have you read to better understand your disease? Was that information given at clinic or found elsewhere?

<u>Comparing to other patients</u>

Is it important to you to understand how your disease compares to other patients with scleroderma?

IF answer is positive, in what ways would you like to know how your disease compares to others (likelihood/rate of progression, mortality, organs affected, etc.)?

IF answer is negative, are you more concerned with your own disease and treatment that is beneficial for you?

How well do you understand your disease compared to other patients with scleroderma?

IF answer is positive, can you describe how your disease compares to other patients with scleroderma?

B.2.2 Provider questionnaire

Interview questions – Providers, Phase 1
For each question, please CIRCLE only ONE answer.

1. How well were you able to assess this patient's current health state?

   Poorly Somewhat Well Very well

2. How well were you able to assess this patient's current scleroderma trajectory?

   Poorly Somewhat Well Very well

3. How long did it take you to review this patient's history and current measurements/reports to assess her (his) disease status and trajectory?
   5-10 minutes 11-15 minutes 16-20 minutes 21+minutes

4. How would you communicate this patient's current disease activity to her (him)?

5. Would you make a change in the plan for testing or treatment today?

   Yes No

   If yes, what changes would you make?

6. If you recommended a change in the plan for testing or treatment, how confident were you in your decision?

   Very uncertain Somewhat uncertain Neither confident nor uncertain Somewhat confident Very confident

7. Why were you confident or not confident?

8. What additional evidence would make you more confident of your decisions?

9. How well do you understand this patient's scleroderma status in comparison to other similar patients?

   Poorly Somewhat Well Very well

10. How well do you understand this patient's scleroderma trajectory in comparison to other similar patients?

    Poorly Somewhat Well Very well

11. How well did your level of understanding of this patient's health status & trajectory adequately support the decisions you had to make?

    Poorly Somewhat Well Very well

12. Which organ systems place this patient at greatest risk of clinical decline in the next year?

13. From the available evidence, how well were you able to surmise this patient's risk of having the following scleroderma-related major clinical events in next year?

    Clinically significant ILD: Poorly Somewhat Well Very well

    Cardiomyopathy: Poorly Somewhat Well Very well

    Pulmonary hypertension: Poorly Somewhat Well Very well

    Scleroderma Renal Crisis: Poorly Somewhat Well Very well

    Ischemic digital ulceration: Poorly Somewhat Well Very well

    Severe GI dysmotility or malnutrition: Poorly Somewhat Well Very well

    Myopathy: Poorly Somewhat Well Very well

    Cancer: Poorly Somewhat Well Very well

Post-visualization questionnaire

1. To what extent did the data visualization help you to review this patient's history and current measurements/reports in order to assess his/her current health state and trajectory?

   Not at all To small extent To a moderate extent To a great extent

2. To what extent would the data visualization help you to explain this patient's disease state and trajectory to her (him)?

   Not at all To small extent To a moderate extent To a great extent

3. Did you gain any new insight into this patient's disease by seeing the data visually presented?

   No Yes

4. Were there some specific parts of the visualizations that helped you gain insight?

   No Yes

   If yes, please specify:

5. Were there some specific parts of the visualization that you found confusing?

   No Yes

   If yes, please specify:

6. Would you suggest any modifications for the visualization?
   No Yes

   If yes, please specify:

7. Is there additional information that would help you better understand your patient's disease?
   No Yes

   If yes, please specify:

B.3 The COMRADE questionnaire

Combined Outcome Measure for Risk Communication and Treatment Decision Making Effectiveness (COMRADE)

These questions ask about how you felt about talking with your doctor **AT THIS VISIT**.

	Strongly Agree	Agree	Neither Agree or Disagree	Disagree	Strongly Disagree
The doctor made me aware of the different treatments available	☐	☐	☐	☐	☐
The doctor gave me the chance to express my opinions about the different treatments available	☐	☐	☐	☐	☐
The doctor gave me the chance to ask for as much information as I needed about the different treatment choices	☐	☐	☐	☐	☐
The doctor gave me enough information about the treatment choices available	☐	☐	☐	☐	☐
The doctor gave enough explanation of the information about treatment choices	☐	☐	☐	☐	☐
The information given to me was easy to understand	☐	☐	☐	☐	☐
I know the advantages of treatment or not having treatment	☐	☐	☐	☐	☐
I know the disadvantages of treatment or not having treatment	☐	☐	☐	☐	☐
The doctor gave me a chance to decide which treatment I thought was best for me	☐	☐	☐	☐	☐
The doctor gave me a chance to be involved in the decisions during the consultation	☐	☐	☐	☐	☐
Overall, I am satisfied with the information I was given	☐	☐	☐	☐	☐
My doctor and I agreed about which treatment (or no treatment) was best for me	☐	☐	☐	☐	☐
I can easily discuss my condition again with my doctor	☐	☐	☐	☐	☐
I am satisfied with the way in which the decision was made in the consultation	☐	☐	☐	☐	☐
I am sure that the decision made was the right one for me personally	☐	☐	☐	☐	☐
I am satisfied that I am adequately informed about the issues important to the decision	☐	☐	☐	☐	☐
It is clear which choice is best for me	☐	☐	☐	☐	☐
I am aware of the treatment choices I have	☐	☐	☐	☐	☐
I feel an informed choice has been made	☐	☐	☐	☐	☐
The decision shows what is most important to me	☐	☐	☐	☐	☐

B.4 Patient questionnaire for Phase 2

Scleroderma health status questionnaire

For each question, please CIRCLE only ONE answer.

1. How well do you understand your scleroderma today?

   Poorly Somewhat Well Very well

2. How well do you understand the trajectory of your scleroderma up to this point?

   Poorly Somewhat Well Very well

3. How well do you understand the likely future trajectory or prognosis of your scleroderma?

   Poorly Somewhat Well Very well

4. How well do you understand your disease compared to most other patients with scleroderma?

   Poorly Somewhat Well Very well

5. Is it important to you to understand how your disease compares to other patients with scleroderma?

   Not at all Somewhat Definitely

6. How well do you understand the reason you have been prescribed or not prescribed certain medications for your scleroderma?

   Poorly Somewhat Well Very well

Post-visualization questionnaire (only for patients randomized to the visualization tool)

1. How well did you understand the meaning of the numbers in the visualizations?

   Not at all Somewhat Very well

2. Were there some specific parts of the visualizations that you liked?

   No Yes

   If yes, please specify:

3. Were there some specific parts of the visualization that confused you?

   No Yes

   If yes, please specify:

4. Would you suggest any modifications for the visualization?
   No Yes

   If yes, please specify:

5. Is there additional information that would better help you understand your disease?
   No Yes

   If yes, please specify:

References

Pattanaik, Debendra, Monica Brown, and Arnold E. Postlethwaite (2011). "Vascular involvement in systemic sclerosis (scleroderma)". In: *Journal of Inflammation Research* 4, pp. 105–125.

Mayes, Maureen D., James V. Lacey, Jennifer Beebe-Dimmer, Brenda W. Gillespie, Brenda Cooper, Timothy J. Laing, and David Schottenfeld (2003). "Prevalence, incidence, survival, and disease characteristics of systemic sclerosis in a large US population". In: *Arthritis and Rheumatism* 48.8, pp. 2246–2255.

Fairweather, DeLisa, Sylvia Frisancho-Kiss, and Noel R. Rose (2008). "Sex Differences in Autoimmune Disease from a Pathological Perspective". In: *The American Journal of Pathology* 173.3, pp. 600–609.

Denton, Christopher P. and Dinesh Khanna (2017). "Systemic sclerosis". In: *The Lancet* 390.10103, pp. 1685–1699.

Steen, Virginia D. (2008). "The many faces of scleroderma". In: *Rheumatic Diseases Clinics of North America* 34.1, pp. 1–15; v.

Shah, Ami A. and Fredrick M. Wigley (2013). "My Approach to the Treatment of Scleroderma". In: *Mayo Clinic Proceedings* 88.4, pp. 377–393.

Allanore, Yannick, Robert Simms, Oliver Distler, Maria Trojanowska, Janet Pope, Christopher P. Denton, and John Varga (2015). "Systemic sclerosis". In: *Nature Reviews Disease Primers* 1.1, pp. 1–21.

Brown, Helen and Robin Prescott (1999). *Applied Mixed Models in Medicine*. Wiley.

Diggle, Peter J., Patrick J. Heagerty, Kung-Yee Liang, and Scott L. Zeger (2002). *Analysis of Longitudinal Data*. Second Edition. Oxford Statistical Science Series.

Harville, David A. (1976). "Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects". In: *The Annals of Statistics* 4.2, pp. 384–395.

Harville, David A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems". In: *Journal of the American Statistical Association* 72.358, pp. 320–338.

Graybill, Franklin A. (1976). *Theory and Application of the Linear Model*. North Scituate, MA: Duxbury Press.

Potthoff, Richard F. and Samarendra N. Roy (1964). "A generalized multivariate analysis of variance model useful especially for growth curve problems". In: *Biometrika* 51.3, pp. 313–326.

Rao, C. Radhakrishna (1965). "The Theory of Least Squares When the Parameters are Stochastic and Its Application to the Analysis of Growth Curves". In: *Biometrika* 52.3, pp. 447–458.

Grizzle, J. E. and D. M. Allen (1969). "Analysis of growth and dose response curves". In: *Biometrics* 25.2, pp. 357–381.

Laird, Nan M. and James Harold Ware (1982). "Random-effects models for longitudinal data." In: *Biometrics* 38, pp. 963–974.

Reinsel, Gregory (1984). "Estimation and Prediction in a Multivariate Random Effects Generalized Linear Model". In: *Journal of the American Statistical Association* 79.386, pp. 406–414.

Shah, Amrik, Nan Laird, and David Schoenfeld (1997). "A Random-Effects Model for Multiple Characteristics With Possibly Missing Data". In: *Journal of the American Statistical Association* 92.438, pp. 775–779.

Sammel, Mary, Xihong Lin, and Louise Ryan (1999). "Multivariate linear mixed models for multiple outcomes". In: *Statistics in Medicine* 18.17, pp. 2479–2492.

Fieuws, Steffen and Geert Verbeke (2004). "Joint modelling of multivariate longitudinal profiles: pitfalls of the random-effects approach". In: *Statistics in Medicine* 23.20, pp. 3093–3104.

Wang, Wan-Lun and Tsai-Hung Fan (2012). "Bayesian analysis of multivariate t linear mixed models using a combination of IBF and Gibbs samplers". In: *Journal of Multivariate Analysis* 105.1, pp. 300–310.

Bloomfield, Peter and Geoffrey S. Watson (1975). "The Inefficiency of Least Squares". In: *Biometrika* 62.1, pp. 121–128.

Tukey, John W. (1948). "Approximate Weights". In: *The Annals of Mathematical Statistics* 19.1, pp. 91–92.

Aitken, Alexander C. (1934). "On Least Squares and Linear Combination of Observations". In: *Proceedings of the Royal Society of Edinburgh* 55, pp. 42–48.

Zellner, Arnold and David S. Huang (1962). "Further Properties of Efficient Estimators for Seemingly Unrelated Regression Equations". In: *International Economic Review* 3.3, pp. 300–313.

Oliveira, Rosa and Armando Teixeira-Pinto (2015). "Analyzing Multiple Outcomes: Is it Really Worth the use of Multivariate Linear Regression?" In: *Journal of Biometrics & Biostatistics* 06.4.

Rubin, Donald B. (1976). "Inference and missing data". In: *Biometrika* 63.3, pp. 581–592.

Zellner, Arnold (1962). "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias". In: *Journal of the American Statistical Association* 57.298, pp. 348–368.

Rebonato, Riccardo and Peter Jäckel (2001). "The Most General Methodology to Create a Valid Correlation Matrix for Risk Management and Option Pricing Purposes". In: *Journal of Risk* 2.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1, pp. 1–48.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team (2019). *nlme: Linear and Nonlinear Mixed Effects Models*.

Hadfield, Jarrod D (2010). "MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package". In: *Journal of Statistical Software* 33.2, pp. 1–22.

Institute of Medicine (2001). *Multiple Sclerosis: Current Status and Strategies for the Future*. Washington, DC: The National Academies Press.

Zeller, Carlos Borelli and Simone Appenzeller (2008). "Cardiovascular Disease in Systemic Lupus Erythematosus: The Role of Traditional and Lupus Related Risk Factors". In: *Current Cardiology Reviews* 4.2, pp. 116–122.

Jain, Samay (2011). "Multi-organ autonomic dysfunction in Parkinson disease". In: *Parkinsonism & related disorders* 17.2, pp. 77–83.

Steen, Virginia D. and Thomas A. Medsger (2000). "Severe organ involvement in systemic sclerosis with diffuse scleroderma". In: *Arthritis and Rheumatism* 43.11, pp. 2437–2444.

Tyndall Anthony J. et al. (2010). "Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database". In: *Annals of the Rheumatic Diseases* 69.10, pp. 1809–1815.

Mcnearney, Terry A., John D. Reveille, Michael Fischbach, Alan W. Friedman, Jeffrey R. Lisse, Niti Goel, Filemon K. Tan, Xiaodong Zhou, Chul Ahn, Carol

A. Feghali-Bostwick, Marvin Fritzler, Frank C. Arnett, and Maureen D. Mayes (2007). "Pulmonary involvement in systemic sclerosis: Associations with genetic, serologic, sociodemographic, and behavioral factors". In: *Arthritis Care & Research* 57.2, pp. 318–326.

Faucett, Cheryl L. and Duncan C. Thomas (1996). "Simultaneously Modelling Censored Survival Data and Repeatedly Measured Covariates: A Gibbs Sampling Approach". In: *Statistics in Medicine* 15.15, pp. 1663–1685.

Wulfsohn, M. S. and A. A. Tsiatis (1997). "A joint model for survival and longitudinal data measured with error". In: *Biometrics* 53.1.

Xu, Jane and Scott L. Zeger (2001). "The Evaluation of Multiple Surrogate Endpoints". In: *Biometrics* 57.1, pp. 81–87.

Rizopoulos, Dimitris and Pulak Ghosh (2011). "A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event". In: *Statistics in Medicine* 30.12, pp. 1366–1380.

Brown, Elizabeth R., Joseph G. Ibrahim, and Victor DeGruttola (2005). "A flexible B-spline model for multiple longitudinal biomarkers and survival". In: *Biometrics* 61.1, pp. 64–73.

Proust-Lima, Cécile and Jeremy M. G. Taylor (2009). "Development and validation of a dynamic prognostic tool for prostate cancer recurrence using repeated measures of posttreatment PSA: a joint modeling approach". In: *Biostatistics (Oxford, England)* 10.3, pp. 535–549.

Garre, Francisca Galindo, Aeilko H. Zwinderman, Ronald B. Geskus, and Yvo W. J. Sijpkens (2008). "A joint latent class changepoint model to improve the prediction of time to graft failure". In: *Journal of the Royal Statistical Society Series A* 171.1, pp. 299–308.

Schoenfeld, Sara R. and Flavia V. Castelino (2015). "Interstitial Lung Disease in Scleroderma". In: *Rheumatic diseases clinics of North America* 41.2, pp. 237–248.

Legendre, Paul and Luc Mouthon (2014). "Pulmonary arterial hypertension associated with connective tissue diseases". In: *Presse Medicale (Paris, France: 1983)* 43.9, pp. 957–969.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson (2020). *shiny: Web Application Framework for R*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

van den Hoogen et al. (2013). "2013 Classification Criteria for Systemic Sclerosis: An American College of Rheumatology/European League Against

Rheumatism Collaborative Initiative: ACR/EULAR Classification Criteria for SSc". In: *Arthritis & Rheumatism* 65.11, pp. 2737–2747.

Brooke, John (1995). "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189.

Edwards, Adrian, Glyn Elwyn, Kerry Hood, Michael Robling, Christine Atwell, Margaret Holmes-Rovner, Paul Kinnersley, Helen Houston, and Ian Russell (2003). "The development of COMRADE—a patient-based outcome measure to evaluate the effectiveness of risk communication and treatment decision making in consultations". In: *Patient Education and Counseling*. Shared decision making in health care 50.3, pp. 311–322.

Muthen, Bengt (2001). "Latent variable mixture modeling". In: *New developments and techniques in structural equation modeling* 2, pp. 1–33.

Ji Soo Kim

jkim478@jhu.edu
(443) 333-6850

Education

Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland
Ph.D. in Biostatistics — Expected October 2020

Rice University, Houston, Texas
B.A in Statistics, Mathematical Economic Analysis, *magna cum laude* — December 2013

Research Experience

Graduate Research Assistant

Johns Hopkins University Department of Biostatistics, Baltimore, Maryland — November 2014 to present

Survival, Longitudinal and Multivariate Models for Health Monitoring

- Designed a Bayesian method for predicting Scleroderma patients' disease state or trajectory given patterns observed in multivariate symptom measures
- Compared and measured relative merits of the model fully utilizing information in multiple longitudinal markers to those of marker-specific models

Development of a Predictive Model of Left Ventricular Failure in Scleroderma

- Identified patients at high risk of left ventricular failure using demographics, clinical features, autoantibody status
- Developed a prediction tool for individual patients' anticipated ejection fraction values at any given time using time-varying ejection fraction data as well as baseline variables

Methods for Evaluating Health Programs with Pre-existing Data

- Built a causal model of child mortality in sub-Saharan African Countries utilizing pre-existing household surveys
- Evaluated district-wise impact of a health program on mortality using mixed effects hierarchical logistic regression models

Assessing Impact of Sleep Stages and Duration on Mortality

- Explored various components of sleep that affect survival for individuals using polysomnography data

Neuroimaging projects using MRI and fMRI images

- Investigated methods of measuring reproducibility of correlation matrices of independent components in fMRI
- Performed longitudinal investigation of multiple sclerosis lesion intensity trajectories in MRI
- Explored methods for faster and more accurate automatic voxel-wise brain segmentation in healthy adult brain MRI

Teaching Experience

Co-instructor

Advanced Biostatistics Topics Seminar Course: Bayesian Hierarchical Models for Individualized Health
Scott L. Zeger and Ji Soo Kim — September 2019

- Designed and taught course materials introducing statistical foundation of Bayesian hierarchical models and computational approaches to fit models for multivariate longitudinal data and make inference on their results

Teaching Assistant

Johns Hopkins University Department of Biostatistics, Baltimore, Maryland

- 140.655: Longitudinal Data Analysis (2020 – Lead TA)
- 140.656: Multilevel Statistical Models in Public Health (2020 – Lead TA)
- 140.653-4: Methods in Biostatistics III & IV (2016 & 2019 – Lead TA)
- 140.641: Survival Analysis I (2018 & 2020)
- 140.622-4: Statistical Methods in Public Health II, III & IV (2018)
- 140.611-2: Statistical Reasoning in Public Health I & II (2015 & 2017)

Consulting Experience

Statistician/Programmer — September 2019 to present

Embold Health, Nashville, Tennessee

- Built Bayesian hierarchical models to measure and compare differences in practices across multiple health providers using multiple binary clinical measures as outcome and made visualizations to help interpret model estimates

Publications and Manuscripts

Peer Reviewed Publications

- Dong-Hoon Choi, Grant Kitchen, **Ji Soo Kim**, Yi Li, Kain Kim, In cheol Jeong, Jane Nguyen, Kerry J. Stewart, Scott L. Zeger & Peter C. Searson, "Two Distinct Types of Sweat Profile in Healthy Subjects While Exercising at Constant Power Output Measured by a Wearable Sweat Sensor." 2019 *Scientific Reports*
- R. Nisha Aurora, Ciprian Crainiceanu, Daniel J Gottlieb, **Ji Soo Kim**, Naresh M. Punjabi, "Obstructive Sleep Apnea During Rapid Eye Movement Sleep and Cardiovascular Disease" 2017 *American Journal of Respiratory and Critical Care Medicine*
- R. Nisha Aurora, **Ji Soo Kim**, Ciprian Crainiceanu, Daniel O'Hearn, Naresh M. Punjabi, "Habitual Sleep Duration and All-Cause Mortality in a General Community Sample." 2016 *SLEEP*
- Jamie Perin, **Ji Soo Kim**, Elizabeth Hazel, Lois Park, Rebecca Heidkamp, Scott L. Zeger, "Hierarchical Statistical Models to Represent and Visualize Survey Evidence for Program Evaluation: iCCM in Malawi." 2016 *PLoS ONE*

In Preparation

- **Ji Soo Kim**, Ami Shah, Laura Hummers, Scott L. Zeger, "Modeling Repeated Multivariate Data to Estimate Individuals' Trajectories with Application to Scleroderma."
- **Ji Soo Kim**, Ami Shah, Laura Hummers, Scott L. Zeger, "Predicting Clinical Events using Bayesian Multivariate Linear Mixed Models with Application to Scleroderma."
- **Ji Soo Kim**, Laura Hummers, John Scott, Samantha Pitts, Lauren Smith, Ayse Gurses, Yushi Yang, Ami Shah & Scott L. Zeger, "Development and Assessment of an Individual Patient Level Data Visualization Tool in Scleroderma to Enhance Clinical Care."

Presentations

Invited Talks

- "Patient Interface Design of Scleroderma Patients' Data", *Johns Hopkins Precision Medicine Centers of Excellence (PMCOE) Directors Meeting, Baltimore, Maryland,* October 2020

- "Modeling Repeated Multivariate Data to Estimate Scleroderma Patients' Current and Future Trajectories", *Hopkins InHealth Precision Medicine Steering Committee Meeting, Baltimore, Maryland,* June 2020
- "Identifying and Optimizing Care for Scleroderma Subgroups", *Johns Hopkins Department of Rheumatology NIH Program Project (P30) Annual Retreat, Baltimore, Maryland*, April 2019
- "Statistical Methods for Program Evaluation: Building Capacity within the Governments of Malawi, Mali, Mozambique, and Tanzania for Improved Evidence-based Decision-making for Health Policy and Program Planning", *National Evaluation Platform Mali Workshop, Dakar, Senegal,* March 2017

Contributed Talk

- "Modeling Repeated Multivariate Data to Estimate Individuals' Trajectories with Application to Scleroderma.", *Eastern North American Region (ENAR) Virtual Meeting,* March 2020

Computing Projects

Developer

Scleroderma Data Visualization Tool (R Shiny App)

- Interactive tool for visualizing patient data and predictions to improve clinicians' assessment of an individual patient's health state and trajectory

Co-developer

inEffect (R Package in development)

- Software for visualizing feature importance and feature effect of machine learning algorithms based upon the relationship between the predictors and outcome of interest to increase the interpretability of machine learning predictions

Honors, Awards, and Activities

JHU Medhacks 2018 Winner of Roivant Sciences - Best Use of Public Healthcare API Award	September 2018
Co-organizer of Biostatistics Computing Club at Johns Hopkins	August 2017 to May 2018
Student Representative for Department of Biostatistics at Johns Hopkins	August 2015 to May 2016
Distinction in Research and Creative Work	Winter 2013
Phi Beta Kappa, Member	Inducted Winter 2013
Louis J. Walsh Scholarship in Engineering	Spring & Fall 2013